China's Paid Trolls and the Incentives of Authoritarian Regimes to Manipulate Information[*]

Jakub Redlicki[†]

May 31, 2015

Abstract

It has been observed in several countries that online commentators are employed by state-controlled institutions to post pro-government opinions on the Internet. We formally analyse the incentives of authoritarian regimes to hire such agents in a model with three types of players: the government, online commentators, and citizens. The model provides a theoretical rationale for the empirical finding of King, Pan and Roberts (2013) that the goal of the Chinese censorship programme is to silence comments that may lead to collective action rather than to suppress criticism of the state. Contrary to intuition, regimes may be more likely to employ online commentators (i) the higher are the citizens' private costs of attacking, (ii) the lower are the portions of the gains from regime change that are accessed only by those who had attacked the regime, and (iii) the higher is the prior mean state of the world. Furthermore, if the citizens are not well-informed about the regime's manipulative practices, the government may have a signal-jamming incentive to employ the commentators even if criticism of the state, rather than the possibility of collective action, is its concern.

Keywords: Information manipulation, Internet, Regime change, Global games, Intrinsic and extrinsic motivation, Signal-jamming. *JEL codes: C72, D74, D83*

1 Introduction

According to the "Freedom on the Net 2013" report by the Freedom House, state-controlled institutions in several countries employ online commentators with the aim of manipulating

[*] I am very grateful to my supervisor, Meg Meyer, for her advice and encouragement. I also thank Avinash Dixit, Paweł Gola, Ian Jewitt, Stephen Morris, and Bartosz Redlicki for helpful comments, as well as seminar participants at the University of Oxford and the RES-York Symposium on Game Theory. I gratefully acknowledge financial support from the Economic and Social Research Council (UK).

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

the information available to citizens.[1] The most well-known is the Chinese case, where the commentators came to be known as the Fifty Cent Party, since they were said to be paid 50 cents (5 mao) for every post that advances the line of the Communist Party of China. With 250,000-300,000 "Fifty Cent Party members", the scale of the online commentator system in China is larger than anywhere else.[2] For this reason, the discussion in this introduction will refer primarily to the Chinese case; however, it is worth noting that the presence of pro-government online commentators has also been documented in countries such as Bahrain, Belarus, Ecuador, Russia, and Sudan.[3]

The principal task of online commentators is to post comments favourable towards party policies and to guide public opinion in that direction. In some cases, they are simply asked to post and reply to threads on a certain forum.[4] However, online commentators may also perform a much broader spectrum of actions, for example, they may monitor, collect, analyse and report online public opinion as well as coordinate with government agencies to provide timely responses and feedback to Internet users.[5] They may also be asked to disrupt online discussions, spread misinformation, and intimidate other Internet users. Notably, online commentators are typically more active when there is a need for online crisis management or state propaganda campaigns, which usually occurs when public discontent is rising.[6]

Some examples of tasks given by superiors have been provided by an anonymous commentator in an interview with Ai Weiwei, a Chinese writer and activist:[7]

"For example, 'Don't spread rumours, don't believe in rumours', or 'Influence public understanding of X event', 'Promote the correct direction of public opinion on XXXX', 'Explain and clarify XX event; avoid the appearance of untrue or illegal remarks', 'For the detrimental social effect created by the recent XX event, focus on guiding the thoughts of netizens in the correct direction of XXXX'."

This also suggests that, although they are monitored by their superiors, online commentators have certain discretion as to the exact methods they use when performing the job.

One of the main challenges of online commentators' work is writing favourable comments about the government in such a way that other Internet users do not immediately notice that

[1] Freedom House, 2013, "Freedom on the Net 2013: A Global Assessment of Internet and Digital Media" (2013), edited by S. Kelly, M. Truong, M. Earp, L. Reed, A. Shahbaz, and A. Greco-Stoner; https://freedomhouse.org/report-types/freedom-net

[2] See King, Pan and Roberts (2013).

[3] Also Cuba, Egypt, Ethiopia, India, Kazakhstan, Kyrgyzstan, Malawi, Malaysia, Mexico, Morocco, Saudi Arabia, South Korea, Thailand, Uzbekistan, Venezuela, and Vietnam. See "Freedom on the Net 2013".

[4] For example, the Hengyang Party School Front Website asked its commentators to "post comments, replies or original commentary threads on the website of Party School Front" (Han, 2012).

[5] This has been reported in Dongjiang District, Ningbo (Han, 2012).

[6] See Han (2012).

[7] Weiwei, A., "China's Paid Trolls: Meet the 50-Cent Party", New Statesman, 17 October 2012; http://www.newstatesman.com/politics/politics/2012/10/china%E2%80%99s-paid-trolls-meet-50-cent-party

they are being paid to do that. As the anonymous commentator testifies in the interview with Ai Weiwei, this requires skill and good understanding of other Internet users' psychology:[8]

> *"You can't write in a very official manner, you must conceal your identity, write articles in many different styles, sometimes even have a dialogue with yourself, argue, debate. (...) The netizens are used to seeing unskilled comments (...). They know what is behind it at a glance. The principle I observe is: don't directly praise the government or criticise negative news. Moreover, the tone of speech, identity and stance of speech must look as if it's an unsuspecting member of public; only then can it resonate with netizens."*

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 are provided with pecuniary incentives. They are often compensated on a piece rate basis (which was an inspiration for 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 the best achievers). The compensation may be higher during unexpected events or if one is guiding public opinion on an important issue with many people posting.[9] The fact that online commentators are paid for posting comments is a crucial element of the model presented in this paper.

The system of employing online commentators has not escaped certain negative implications. It has been noted that the system may increase the distrust of Internet users in any favourable opinion about the state since, when online commentators are present, any pro-government voice can be suspected to be paid propaganda. This effect, which is a crucial feature of the model presented in this paper, has been well summarised by P. Link:[10]

> *"Posts [for pay] also run the risk of undermining opinion that might be genuinely pro-government, because they make any pro-government comment subject to the suspicion that it was done for money. In some circles, mockery aimed at fifty-centers has expanded to include regime apologists of any kind. Someone who thinks that External Propaganda might actually be doing some good by watching the Internet is called a 'self-employed fifty-center'. Westerners who praise the CCP [the Chinese Communist Party] are 'foreign fifty-centers'."*

Most crucially, the model presented in this paper is closely connected to the empirical findings of King, Pan and Roberts (2013), who have proposed two distinct theories of what constitutes the goals of the Chinese regime as implemented in their censorship programme:

[8] Ibid.

[9] See Han (2012).

[10] Link, P., "Censoring the News Before It Happens", The New York Review of Books Blog, 10 July 2013; http://www.nybooks.com/blogs/nyrblog/2013/jul/10/censoring-news-before-happens-china/

the "state critique theory" and the "theory of collective action potential". According to the former, censorship in China is aimed at restricting any criticism of the government and its policies. The latter assumes instead that the goal 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.

With theories defined in that way, it could be that either or both are correct or incorrect. In order to check whether any of these two theories is correct, King et al. (2013) devised a system which allowed them to analyse the content of millions of social media posts originating from social media providers all over China before the Chinese government was able to evaluate posts that were to be censored. Then, using modern computer-assisted text analytic methods, they compared the content of posts censored to those not censored over time in several topic areas. They provide empirical evidence that, with only few exceptions, the state critique theory is incorrect, while the theory of collective action potential is correct. This means, in other words, that the censorship programme is aimed at curtailing collective action by silencing comments that could lead to or reinforce social mobilization, regardless of the comments' content.

We extrapolate the empirical findings of King et al. (2013) to the programme of employing online commentators, which is arguably a more subtle form of manipulating information than censorship is. Nevertheless, it serves a similar purpose, which is to keep in control the information accessed by the Chinese citizens, and therefore we argue that such extrapolation is valid. The results of our model relate directly to the conclusion of King et al. that the aim of the government's programme is to stop the spread of information that could lead the citizens to a collective action. The following section outlines the structure of the paper.

1.1 Outline of the Analysis

The aim of this paper is to formally analyse the incentives of regimes to employ online commentators. We are interested in how these incentives are affected by different features of the environment and we aim to pin down those which determine the value of the payment for the commentators. We analyse the regime's incentives in two different theoretical settings, and we argue that they illustrate the two theories of King et al. (2013): the state critique theory and the theory of collective action potential.

Section 2 outlines the setup of the model, that is, the common background of both theoretical settings analysed in the paper. In particular, in both environments, there are three types of economic agents: the government, online commentators, and citizens. Furthermore, the government's utility function consists of a benefit function less the direct cost of employing the online commentators, with a parameter introduced to indicate the relative importance of the benefit component in the utility function. However, the two settings differ with respect to the actions taken by the citizens based on their posteriors, and with respect to the benefit functions of the government.

In the first setting, which is discussed in Section 3, the government's benefit function is linear in the citizens' posteriors about the state of the world. We will refer to it as the "linear model" and we will argue that it is an economic interpretation of the state critique theory of King et al. In this environment, we show that the government has no incentives to employ online commentators as long as the citizens are sophisticated enough to know the value of the online commentators' payment.

In the second setting, which is analysed in Section 4, the citizens decide—based on their posteriors—whether to attack the regime or not, and the government's benefit function is the ex ante probability that the attack is unsuccessful. We will refer to it as the "threshold model". We start the analysis with a benchmark setting in which there are no coordination issues between the citizens. The purpose of this benchmark is to provide better intuition for the transition between the linear model of Section 3 and the global game of regime change of Section 4. Coordination between the citizens is at the heart of this game, and we will argue that this setting illustrates the theory of collective action potential, as defined by King et al.

Furthermore, we demonstrate that the incentives of regimes to employ online commentators are non-monotonic in (i) the citizens' private costs of attacking the regime, (ii) the portions of the gains from regime change that are accessed only by those who had attacked the initial regime, and (iii) the prior mean state of the world, which could be interpreted as the general state of the country. The regimes are most likely to employ online commentators when these parameters take moderate values, which suggests that, in fact, we may observe the system of online commentators become more widespread as the social and economic situation improves or as the costs of protesting increase.

Section 5 reconsiders the results of the model presented in Section 3 in a setting where the online commentators' payment is no longer common knowledge. It turns out that, if the payment is not observed by the citizens, the government may have a signal-jamming incentive to employ online commentators. As a result, the government may find it optimal to hire them even if it cares solely about state critique (rather than the possibility of a collective action). This indicates that one potentially crucial factor in the empirical results of King et al. (2013) is how well-informed the citizens are about the regime's manipulative practices.

Section 6 contains further discussion, ideas for extensions and concluding remarks. Section 7 contains the appendix.

1.2 Related Literature

The preferences of online commentators, which form a core part of the setup of our model, are inspired by the preferences of agents in Bénabou and Tirole (2006). In this paper, Bénabou and Tirole study the behaviour of agents who choose the extent of their participation in some prosocial activity, e.g. contribution to a public good. The direct benefit of an agent from a given level of participation is a function of the agent's intrinsic valuations for money and for contributing to the social good. In our model, the number of positive comments that an online commentator writes about the government corresponds to Bénabou and Tirole's level of participation in a prosocial activity. They show that rewards or punishment can create doubt about the true motive for which good deeds are performed, which could lead to crowding out of prosocial behaviour by extrinsic incentives (the so-called "overjustification effect"). As a result, small rewards and punishments can be counterproductive as incentive mechanisms.[11]

However, since online commentators are supposed not to reveal their Internet identities to the real world, we assume that they do not have reputational incentives, and so—unlike the agents in Bénabou and Tirole—they do not feel the need to appear as prosocial and not greedy. Therefore, the social signalling component is absent in their utility function. For Bénabou and Tirole, reputational incentives are essential for the formal representation of counterproductive monetary incentives, while we are more interested in the signal extraction problem *per se*.[12]

The threshold model presented in Section 4 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). Since Morris and Shin (2003) provide an excellent survey of the global games literature, we will only focus on papers that are most relevant to ours.

Our model is particularly closely related to Edmond's (2013) model of information manipulation and political regime change, which builds on the global games literature. In his model, like in ours, the regime can be overthrown if enough citizens participate in an uprising. A further similarity is that the citizens are imperfectly informed about the regime's ability to resist an uprising and the regime can engage in propaganda that, taken at face value, makes the regime appear stronger than it really is.

However, there are also important differences between Edmond's paper and ours. In his model, the regime is informed about the state of the world and takes a type-specific hidden action that cannot be directly observed by the individual citizens receiving the information. Edmond shows that the regime's information manipulation, which is a form of signal-jamming in his model, can be effective in equilibrium. By contrast, in our model, the regime is uninformed (i.e. it does not know the state of the world before signing contracts with the commentators) and its action is not hidden (i.e. the citizens can observe the payment given

[11] For example, compare with Titmuss (1970), who argued that paying blood donors could reduce supply, and with Gneezy and Rustichini (2000), who found that fining parents for collecting their children late from day-care centres led to more late arrivals.

[12] The interplay between intrinsic and extrinsic motivation is also discussed at length in Bénabou and Tirole (2003), however, its mechanisms are different in that paper. The authors analyse how performance incentives offered by an informed principal (e.g., a manager) can adversely impact an agent's (e.g., an employee's) perception of the task, or of his own abilities. They show that incentives are only weak reinforcers in the short run, and negative reinforcers in the long run.

to the commentators). Moreover, while in Edmond's paper the regime's manipulation affects only the level of the apparent state of the world, in our model it also affects the precision of the private signal.

Our paper discusses the design of information disclosure policies, which is also the subject of Metz (2002) and Heinemann and Bannier (2005). Both of these papers analyse the implications of currency crises in a global games setting. More precisely, they consider the central bank's (or the government's) optimal rules for information dissemination that would minimise the probability of currency crises. They show that full transparency is not always optimal: when the prior beliefs about economic performance are good, the central bank should disclose imprecise information. The comparative statics analysis in our threshold model with coordination is closely related to the results of these two papers. This similarity is further discussed in Section 4.

The potential impact of well-informed individuals on coordination between followers is also investigated by Loeper, Steiner and Stewart (2013). They present a coordination game in which followers observe the decision choices of experts whose interests may not coincide with those of followers. It is shown that the choices of the opinion leaders can have a large effect on outcomes even though the followers are assumed to be Bayesian decision-makers who know the distribution of experts' biases and each expert can influence only a negligible share of the population. Angeletos, Hellwig and Pavan (2007), Boix and Svolik (2013), Bueno de Mesquita (2010), Chassang and Padró-i-Miquel (2010), and Shadmehr and Bernhardt (2011) provide other examples formal analyses and political economy applications of global games.

The linear model in which the online commentators' payment is not observed by citizens (see Section 5) is closely related to models of managerial career concerns, e.g., Holmström (1999), and Dewatripont, Jewitt and Tirole (1999a, 1999b). The state of the world in our model corresponds to the manager's talent in the models of career concerns, while the online commentators' payment is comparable to the manager's effort level. Further discussion of the relationship of our model with economics of career concerns is provided in the second part of the Appendix.

Finally, it is also worth noting that a phenomenon similar to employing online commentators by authoritarian regimes is fairly common among private companies.[13,14] Dellarocas (2010) offers a theoretical analysis of how firms manipulate consumer perceptions by posting costly anonymous messages that praise their products. In particular, he investigates the impact of such behavior on firm profits and consumer surplus. On the empirical front, Mayzlin, Dover and Chevalier (2014) analyse the differences in reviews for a given hotel between two websites: in one of them only a customer could post a review (Expedia.com), while in the

[13] http://www.ag.ny.gov/press-release/ag-schneiderman-announces-agreement-19-companies-stop-writing-fake-online-reviews-and

[14] http://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html

other anyone could post (TripAdvisor.com). Based on this, they determine the characteristics of firms that are most likely to order fake reviews. Nevertheless, although the incentives of firms to employ paid online commentators and reviewers are certainly an idea that is worth exploring, we focus on the political economy application to authoritarian regimes.

2 The Model

There are three types of economic agents: the government, online commentators, and citizens.

Government. The government, who is the principal in this model, hires commentators to praise the government on the Internet and pays them a wage of $y \geq 0$ per each positive comment they write. Two different objective functions of the government will be analysed.

Commentators. An online commentator's intrinsic valuation for money is denoted by v, and the state of the world (which is assumed to be identical with the strength of the regime) is given by θ. The commentator's utility from praising the government at a level $a \in R$ is given by:

$$U_A := (\theta + vy)a - C(a)$$

The praise level, a, is interpreted in the model as the number of pro-government comments written by an online commentator. Intuitively, the better is the state of the world, θ, the more eager is a commentator to write positive comments about the governement. This may also reflect the fact that it is easier for the commentator to praise the government when the state of the world is good. But the degree to which he praises the government depends also on the payment y, which he receives from the government. The extent to which a commentator is responsive to this payment depends on his intrinsic valuation for money, v.[15]

However, the commentators also face a cost of writing posts, $C(a)$, which can be understood as a cost of effort. The cost function is assumed to be convex and, for more tractability, we assume that it has the form $C(a) = \frac{1}{2}ka^2$, where k is a parameter that measures 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 $a^* = (\theta + vy)/k$. We will assume that $k = 1$ to make the exposition easier.

Citizens. The citizens observe a noisy measure of the online commentator's action, $x := a + \varepsilon$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, and use it to update their information about the true state of the world, θ. This adds another source of noise to the citizen's information about θ—besides the noise generated by the variability in the commentators' intrinsic valuation for money, v, which

[15] As it has already been stated and discussed in Section 1.2, these preferences of online commentators are inspired by the preferences of agents in Bénabou and Tirole (2006) for prosocial behaviour.

disappears completely when the payment, y, is zero. One interpretation for the noise term ε is that the citizen observes only a sample of the commentator's posts and the noise captures the (inevitable) mistakes in the citizen's predictions about how many posts the commentator has written in total.

Informational structure. It is assumed that only the online commentators observe the exact values of θ and v; the former is realised after the contract is signed, and the latter is the commentator's private knowledge. However, the distribution of θ and v is common knowledge and so it is known by the government as well as the citizens. The variables θ and v are jointly normally distributed with:

$$\begin{pmatrix}\theta \\ v\end{pmatrix} \sim N\left(\begin{matrix}\overline{\theta} \\ \overline{v}\end{matrix} \; , \; \begin{bmatrix}\sigma_\theta^2 & \sigma_{\theta v} \\ \sigma_{\theta v} & \sigma_v^2\end{bmatrix}\right)$$

The assumption of $\overline{v} > 0$ implies that most online commentators respond to the payment $y > 0$ by increasing the number of posts written. We additionally assume that $\overline{\theta} > 0$, which ensures that, on average, the commentators write a strictly positive number of posts, and therefore the government's expected cost of employing online commentators is strictly positive.[16] Furthermore, we assume that $\sigma_{\theta v} = 0$, which means that there is no correlation between the state of the world and the agent's intrinsic valuation for money.[17]

Finally, we assume initially that the value of the online commentators' payment, y, is observed by the citizens, however, this assumption is later relaxed in the linear model.

Timing of the game. The game proceeds in the following steps:

1. The government offers a contract to the commentator: a payment of y per positive comment.

2. The state of the world, θ, is realised and the commentator learns it. Based on that and on his intrinsic valuation for money, v, the commentator decides how many positive comments to write, i.e. he chooses his action a.

[16] Normality yields here greater tractability at the cost of allowing the level of praise for the government—interpreted in the model as the number of posts written by an online commentator—to take implausible negative values, which would result in the government being paid by the commentator rather than the other way round. This is because, with normal distributions of θ and v, it is absolutely possible that both θ and v are so low that the commentator's optimal choice of a is negative. A similar problem has been observed by Bénabou and Tirole (2006), however, they also noted that one can make the probability of such realizations arbitrarily small by choosing the relevant means large enough. Nevertheless, the joint normal distribution of θ and v stated above should be interpreted as a local approximation.

[17] It could be argued that allowing for a positive correlation between θ and v would be more realistic. One explanation for this is that, when the state of the world is good, one could imagine that online commentators would feel less guilt about being responsive to the payment that they receive from the government, and thus their intrinsic valuation for money would also be more likely to be high. Nevertheless, in order to make the analysis more tractable, we assume that this correlation is zero.

3. The citizen observes a noisy private signal of the commentator's action, $x := a + \varepsilon = \theta + vy + \varepsilon$.

4. Given the observed private signal, x, and the knowledge of how much the commentators are paid per positive comment, y, the citizen forms a posterior expectation about the state of the world, θ (and possibly takes an action based on it).

5. The government's payoffs are realised.

3 The Government Fears State Critique

Suppose that the government fears state critique and that, for this reason, it would like the average public opinion to be as high as possible. Thus, it is hurt by a deterioration in a citizen's opinion regardless of how good or bad their initial stance regarding the government was. At the same time, the government faces a direct cost of employing the online commentators, which comes from the fact that they need to be paid for the comments they write.

This setting can be illustrated with a utility function of the government, in which the benefit function is linear in the citizen's posterior about the state of the world and there is a parameter μ to denote the importance of the benefit function relative to the direct cost of employing the online commentators. Mathematically, the government's utility can be expressed as

$$U_P(y;\cdot) \quad := \quad \mathbb{E}_{\theta,v,\varepsilon}\left[\mu\mathbb{E}\left[\theta \mid x, y\right] - ay\right], \tag{1}$$

where the notation $U_P\left(y;\cdot\right)$ is used to underline the fact that the government's utility is a function of the online commentators' payment, y, and a number of exogenous parameters. As there is no problem of coordination among citizens, we assume here—without loss of generality—that there is only one commentator and only one citizen.

The government's utility function stated in (1) provides an economic interpretation of what constitutes the objectives of the government according to the *state critique theory*, defined by King et al. (2013):

> *"(...) [S]tate 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 government fears state critique, its utility should change proportionally with the citizens' beliefs about the state of the world. Consequently, from the government's point of view, small changes in the citizens' 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 citizens' initial opinions. By contrast, if the government cared about the possibility of collective action, then given that citizens' actions can change more radically in response to small changes in their opinions, the government's utility function would have to exhibit some non-linearity with respect to the citizens' posteriors. This possibility will be explored in the subsequent section, in which the citizens' actions are step functions with respect to their posteriors.

Applying the standard results of the normal learning model, we note that the posterior expectation of the state of the world, θ, conditional on a given private signal, x, is

$$\mathbb{E}\left[\theta \mid x\right] = \bar{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\left(x - (\bar{\theta} + \bar{v}y)\right). \tag{2}$$

Thus, the higher is the online commentators' payment, y, the less informative is the private signal about θ, which makes the citizen put more weight on the prior mean state of the world, $\bar{\theta}$, and less on the private signal, x, when forming her posterior.

Proposition 1. *If the government's benefit is linear in the citizens' posterior expectations and the online commentators' payment, y, is observed by citizens, the government's optimal choice of the payment is $y^* = 0$.*

Proof. Using the results of the normal learning model, we can rewrite the government's utility function stated in (1) as:

$$\begin{aligned} U_P(y;\cdot) &= \mathbb{E}_{\theta,v,\varepsilon}\left[\mu\left(\bar{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\left(x - (\bar{\theta} + \bar{v}y)\right)\right) - ay\right] \\ &= \mathbb{E}_{\theta,v,\varepsilon}\left[\mu\left(\bar{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\left(\theta + vy + \varepsilon - (\bar{\theta} + \bar{v}y)\right)\right)\right] - (\bar{\theta} + \bar{v}y)y. \end{aligned} \tag{3}$$

By the law of iterated expectations, this simplifies to $U_P(y;\cdot) = \mu\bar{\theta} - (\bar{\theta} + \bar{v}y)y$. Regardless of the value of the payment y chosen by the government, the ex ante expectation of the citizen's posterior is always equal to the mean state of the world, $\bar{\theta}$. As a result, the government's marginal benefit from increasing y is zero for all values of y. Since the government also faces a direct cost of increasing the payment y and this cost is increasing in y, it follows that the optimal payment for the online commentators is here $y^* = 0$. □

Hence, if the government fears state critique in the sense of having the utility function stated in (1), then it has no incentives to employ online commentators. The fact that the online commentators need to be paid for the posts they write is crucial here: if the government did not care about the direct cost of paying online commentators, then any value of the commentators' payment, y, would do the job equally well.

Proposition 1 provides some theoretical rationale for the empirical findings of King et al. (2013) regarding the objectives of the Chinese government's programme of censorship and information control. It implies that, if the regime's objective were to maximise the ex ante expectation of the citizens' posteriors less the direct cost of employing online commentators, then the government would have no incentives to manipulate the information. Thus, the model suggests that, if the government is employing paid online commentators in the setup described in Section 2, then it cannot be due to the government being concerned about the average public opinion, and therefore the government's objective must be different.

It is important to note here, however, that the result in Proposition 1 relies on the assumption that the online commentators' payment, y, is observed by the citizens. That is, the Internet users need to be fairly sophisticated and well-informed about the regime's practices for the government to have no incentive to employ online commentators. The possibility that the payment is not observed by the citizens will be discussed in Section 5.

4 The Government Fears Collective Action

In this section, we analyse a model with the following differences relative to the model of state critique: (i) the citizens, based on their posteriors, decide in Stage 4 of the game whether to attack the status quo or not, (ii) their attack is successful only if a sufficient number of them participates, and (iii) the government's benefit function is the ex ante probability that the attack is unsuccessful. In other words, the government no longer cares about the citizens' opinions *per se*, but only does so to the extent that they may lead to a collective action that will overthrow the regime. Because there is a threshold that determines whether the regime survives or not, this variant of the model will be referred to as the "threshold model".

We now introduce a few additional assumptions into the model so as to accommodate for the possibility of a coordinated attack on the regime by citizens.

4.1 Preliminaries

Commentator-citizen relationship. There is a continuum of commentators of measure 1, indexed by i and uniformly distributed over $[0, 1]$. In addition to this, there is a continuum of citizens of measure 1, who are uniformly distributed over $[0, 1]$ and are also indexed by i: it is assumed that there is a one-to-one correspondence (bijective function) between commentators and citizens. This means that each citizen follows only one commentator and each commentator is listened by only one citizen. The utility function of the commentators and the assumptions about the distribution of the parameters remain the same and, as before, we assume that the online commentators' payment y is observed by the citizens.

Citizens' payoffs. In Stage 4 of the game, the citizens simultaneously choose between two actions: they can either attack the status quo (e.g. protest, join a revolt, etc.) or refrain from attacking. The regime is overthrown if and only if the measure of citizens attacking, which we denote by D, is no less than the state of the world, $\theta \in \mathbb{R}$, which is assumed here to be identical with the strength of the regime.[18] A citizen's incentive to attack thus increases with the aggregate size of the attack, implying that the citizens' action choices are strategic complements.

The citizens' payoffs from the two possible actions depend on whether the attack turns out to be successful:

	Attack successful ($D \geq \theta$)	Status quo remains ($D < \theta$)
Attack	$1 - c$	$-c$
Do Not Attack	$1 - \delta$	0

The payoffs of the citizens are realised in Stage 5 of the game, that is, once it becomes known whether the regime has survived or not.

Intuition behind parameters c and δ. Parameter $c \in (0, 1)$ denotes the relative cost of attacking, whereas parameter $\delta \in (c, 1]$ reflects the fact that, once the regime is overthrown, those who had attacked it gain privileged status relative to those who had not.[19] One could understand δ as the portion of the payoffs from regime change that can be appropriated only by those who participated in the protest or revolution.

The assumption $\delta > c$ ensures that the game does not collapse to a prisoner's dilemma: if it were that $\delta < c$, then not attacking would be a strictly dominant strategy. In other words, for the revolution to occur with positive probability, the portion of the payoffs which only the attackers can access must be high enough.

The assumption that $\delta \leq 1$ ensures that non-attackers are not worse off after the regime change. Nevertheless, one could imagine a situation where $\delta > 1$ and the citizens who do not participate are worse off after the regime change, for example, because of prosecution by the new regime, ostracism leading to diminished career prospects, etc. However, we assume that $\delta \leq 1$ so that δ could be interpreted as the portion of the payoffs accessed only by the attackers.

[18] In fact, this assumption could be relaxed so that the strength of the regime is a function of the state of the world, say $s(\theta)$, which is strictly increasing in θ (but is not necessarily a linear function θ). Qualitatively, the main results of the model would be preserved.

[19] The idea of a privileged status of attackers appears in political science literature, e.g. Bueno de Mesquita (2010).

It is also worth noting that, as long as $\delta < 1$, there are positive externalities of attacking: if a citizen attacks the current regime and the attack turns out to be successful, the non-attackers are also better off.

Heterogeneity of information about the state of the world. Heterogeneity of the citizens' information about the state of the world, θ, is an important component here. If θ were commonly known by all agents, then we would have three cases to consider from which few conclusions could be drawn. If $\theta > 1$, every player would have a dominant strategy not to attack. If $\theta \in (0, 1]$, there would exist two pure-strategy equilibria, one in which all agents attack and the status quo is abandoned ($D = 1 \geq \theta$), and another in which no agent attacks and the status quo is maintained ($D = 0 < \theta$). Finally, if $\theta \leq 0$, every player would have a dominant strategy to attack.

In our model, the citizens' information about θ is heterogeneous, which is driven by both (i) the heterogeneity of the commentators' intrinsic valuation for money, v, and (ii) the random noise, ε. This heterogeneity allows us to apply the global games framework, which was introduced by Carlsson and van Damme (1993) and further developed by Morris and Shin (1998, 2001). As Morris and Shin (1998) have shown in a model of self-fulfilling currency attacks, if θ is not common knowledge, there is a unique equilibrium as long as agents receive sufficiently precise information on θ. Thus, besides being a more natural assumption about the information structure in the society, heterogeneity of the citizens' information about θ also alleviates the problem of multiplicity of equilibria and allows for considerably tighter predictions.

Government's objective function. The government aims to minimise the probability of regime change, while also taking into account the direct costs of paying the online commentators for the posts they write. Mathematically, the government's utility function can be expressed as

$$U_P(y; \cdot) := \mu Prob(no\, regime\, change) - \mathbb{E}\left[a\right] y, \tag{4}$$

where μ is a weighting parameter denoting the relative importance of minimising the probability of a regime change. As before, the first term, which is here equal to the probability of no regime change multiplied by the weighting parameter, will be referred to as the government's benefit.

Crucially, the government's utility depends on the citizens' posteriors about the state of the world only to the extent that they affect the probability that the regime is overthrown. This can happen in the threshold model only via a collective action of citizens and, therefore, this setting provides an economic interpretation for the *theory of collective action potential*, as defined by King et al. (2013):

"[According to the theory of collective action potential,] the target of censorship is people who join together to express themselves collectively, stimulated by someone other than the government, and seem to have the potential to generate collective action."

4.2 The Perfect Coordination Benchmark

In order to better explain the transition from the linear to the threshold model, we first consider a benchmark with no coordination issues between citizens.

We simplify here the setting presented in Section 4.1 by assuming that if the state of the world, θ, is less than an exogenous critical state, θ^*, the citizen's attack is guaranteed to be successful. By doing this, we are assuming away any coordination problems between the citizens, and therefore we can return—without loss of generality—to the setting with only one citizen and only one commentator.

The citizen's payoffs remain the same as before, however, one comment is necessary. In order to bring this model closer to the general case with coordination issues among citizens, we now assume that, even if the citizen does not attack, there is some *deus ex machina* probability that the regime is under attack. If that attack is successful, the citizen receives a payoff of $1-\delta$, where parameter δ reflects the fact that, once the regime is overthrown, her payoff is higher if she had been one of the attackers.

Given an observed private signal, x, the citizen's expected payoff from attacking is $Prob\,(attack\,successful \mid x) - c$, while her expected payoff from choosing not to attack is $(1-\delta)\,Prob\,(attack\,successful \mid x)$. Thus, the citizen attacks if and only if

$$Prob(attack\,successful \mid x) \geq \frac{c}{\delta}, \tag{5}$$

and she does not attack otherwise. Fraction $\frac{c}{\delta}$ could be interpreted here as a measure of how strong the citizen's posterior belief must be for her to be willing to attack. If $\frac{c}{\delta} = \frac{1}{2}$, then the citizen's position regarding the attack is neutral in the following sense: she attacks if the probability that the attack is successful is greater than the probability that it is not.[20]

Lemma 1. *Consider the threshold model with no coordination issues between citizens, where the online commentators' payment, y, is publicly observed. Given the citizen's optimal choice of whether or not to attack, the government's ex ante utility is:*

$$U_P(y;\cdot) = \mu\Phi\left(\left(\bar{\theta}-\theta^*\right)\frac{\sqrt{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2} + \sqrt{\frac{y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\,\Phi^{-1}\left(\frac{c}{\delta}\right)\right) - \left(\bar{\theta}+\bar{v}y\right)y. \tag{6}$$

Proof. See the Appendix. □

[20] This special case has certain interesting features which will be discussed later in more detail.

Lemma 1 shows the government's ex ante utility when the citizen optimally chooses whether to attack the regime or not. The main implications of this result are in line with common intuition. Keeping the online commentator's payment fixed, the government's benefit is increasing in the prior mean state of the world, $\bar{\theta}$, and is decreasing in the exogenous critical state, θ^*. This occurs because the higher is the former relative to the latter, the more likely it is that the state of the world, θ, will turn out to be higher than the critical state—which would imply that the citizen's attack on the regime is not successful.

Moreover, the government's benefit is increasing in the citizen's cost of attacking, c, and is decreasing in the portion of the gains from regime change that is accessed only by the attackers, δ. The intuition here is that as the cost of attacking becomes higher, the citizen chooses to attack the regime only if her private signal, x, is very low. Consequently, it becomes less likely that she observes such a signal, and as a result the regime is less likely to be attacked and ultimately overthrown. It follows that the government's benefit function becomes larger. The mechanism behind a decrease in the portion of the gains from regime change that is accessed only by the attackers, δ, is equivalent.

However, from the perspective of the government's choice of the online commentator's payment, it is important to look at what happens at the margin. The first order condition of the government's utility maximisation problem yields:

$$\mu \frac{y\sigma_v^2}{\sqrt{\sigma_\theta^2}} \phi(\cdot) \left[\left(\bar{\theta} - \theta^* \right) \frac{1}{\sqrt{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\sqrt{\sigma_\theta^2}} + \frac{1}{\sqrt{y^2\sigma_v^2 + \sigma_\varepsilon^2}} \Phi^{-1}\left(\frac{c}{\delta}\right) \right] \quad = \quad \bar{\theta} + 2\bar{v}y. \quad (7)$$

The left-hand side of (7) can be interpreted as the government's marginal benefit of increasing the online commentator's payment, whereas the right-hand side can be seen as the respective marginal cost. One can make the following observations about the marginal benefit:

Lemma 2. *Consider the threshold model with no coordination issues between citizens. The government's marginal benefit from increasing the payment:*

(i) is equal to zero when the payment is zero and/or there is no uncertainty about the citizen's intrinsic valuation for money, σ_v^2;

(ii) is strictly positive if and only if $\theta^* < \bar{\theta} + \sqrt{\frac{\sigma_\theta^2\left(\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2\right)}{y^2\sigma_v^2 + \sigma_\varepsilon^2}}\Phi^{-1}\left(\frac{c}{\delta}\right)$ and the conditions in part (i) do not hold, and it is negative otherwise;

(iii) approaches zero as the payment becomes arbitrarily large;

(iv) approaches zero for all values of y when at least one of $\bar{\theta}$ and $\frac{c}{\delta}$ becomes arbitrarily large.

Proof. Parts (i) and (ii) are immediately clear from (7).

For part (iii), analyse

$$\phi\left(\cdot\right)=\phi\left(\left(\bar{\theta}-\theta^{*}\right)\left(\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)^{\frac{1}{2}}\left(\sigma_{\theta}^{2}\right)^{-1}+\left(y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)^{\frac{1}{2}}\left(\sigma_{\theta}^{2}\right)^{-\frac{1}{2}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right). \tag{8}$$

If $\theta^{*}\neq\bar{\theta}+\sqrt{\sigma_{\theta}^{2}}\Phi^{-1}\left(\frac{c}{\delta}\right)$, we have that $\lim_{y\to\infty}\phi\left(\cdot\right)=\left[\phi\left(\pm\infty\right)\right]=0$, while the remaining terms in the expression for the marginal benefit approach some constant as $y\to\infty$. Hence, the marginal benefit approaches zero. On the other hand, if $\theta^{*}=\bar{\theta}+\sqrt{\sigma_{\theta}^{2}}\Phi^{-1}\left(\frac{c}{\delta}\right)$, then $\lim_{y\to\infty}\phi\left(\cdot\right)=k$, where $k\neq0$. But then it is also the case that

$$\lim_{y\to\infty}\left[\left(\bar{\theta}-\theta^{*}\right)\left[\sigma_{\theta}^{2}\left(\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)\right]^{-\frac{1}{2}}+\left(y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)^{-\frac{1}{2}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right]=0,$$

which completes the proof of part (iii).
For part (iv), note that:

$$\lim_{\bar{\theta}\to\infty}\frac{\partial\mu\Phi\left(\cdot\right)}{\partial y}=\lim_{\bar{\theta}\to\infty}\mu\frac{y\sigma_{v}^{2}}{\sqrt{\sigma_{\theta}^{2}}}\phi(\cdot)\left[\left(\bar{\theta}-\theta^{*}\right)\frac{1}{\sqrt{\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}\sqrt{\sigma_{\theta}^{2}}}+\frac{1}{\sqrt{y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right]=0,$$

where $\Phi\left(\cdot\right)=\Phi\left(\left(\bar{\theta}-\theta^{*}\right)\frac{\sqrt{\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}}{\sigma_{\theta}^{2}}+\sqrt{\frac{y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}{\sigma_{\theta}^{2}}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right)$ and the last equality follows from two properties of the normal distribution, $\phi'\left(x\right)=x\phi\left(x\right)$ and $\lim_{x\to\pm\infty}\phi'\left(x\right)=0$. The argument for $\frac{c}{\delta}$ follows the same lines. □

Part (i) of Lemma 2 implies that when there is no variance in the citizen's intrinsic valuation for money, $\sigma_{v}^{2}=0$, then the government's marginal benefit of increasing the online commentator's payment, y, is zero for any given value of y. Thus, some degree of uncertainty about the citizen's intrinsic valuation for money is necessary to obtain the result that the government may choose a strictly positive value of the online commentator's payment, $y^{*}>0$. The fact that the marginal benefit also equals zero when the payment is zero is closely related: in that case, the variation in the commentator's intrinsic valuation for money does not lead to any additional noise in the citizen's private signal.

The main implication of part (ii) of Lemma 2 is that, in the threshold model, the government's marginal benefit of increasing the payment, y, may be strictly positive, and hence the government may find it optimal to employ online commentators. This contrasts with the linear model presented in Section 3, where the government's marginal benefit was always zero, and hence the government never had an incentive to employ them. The following definition will prove useful for subsequent analysis:

Definition 1. *The threshold of positive marginal benefit, $\tilde{\theta}\left(y\right)$, is a function of the online commentators' payment such that the government's marginal benefit from increasing the payment is strictly positive (negative) if and only if $\theta^{*}\left(y;\cdot\right)<\left(>\right)\tilde{\theta}\left(y\right)$.*

As can be seen from part (ii) of Lemma 2, the threshold of positive marginal benefit in the benchmark model is given by

$$\tilde{\theta}\left(y\right)=\bar{\theta}+\sqrt{\frac{\sigma_{\theta}^{2}\left(\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)}{y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}}\Phi^{-1}\left(\frac{c}{\delta}\right). \tag{9}$$

The second interesting observation about part (ii) of Lemma 2 is that the stated result contrasts with common intuition. It implies that the government's marginal benefit from increasing the payment is more likely to be strictly positive when (i) the mean state of the world, $\bar{\theta}$, becomes larger, (ii) the citizen's cost of attacking, c, becomes higher, and (iii) the portion of the payoffs accessed only by those who had attacked the regime, δ, becomes smaller. More intuition for these results is provided in the discussion of Propositions 2 and 3.

Part (iii) of Lemma 2 implies that as long as there is—for all levels of online commentator's payment—some strictly positive marginal cost of increasing it, the optimal payment chosen by the government will always be finite.

Finally, the intuition behind part (iv) of Lemma 2 is that as the mean state of the world, $\bar{\theta}$, or the citizen's relative cost of attacking, $\frac{c}{\delta}$, becomes arbitrarily large, the probability of the regime being successfully overthrown becomes very small, and therefore the government's additional benefit from increasing the payment for online commentators fades and then virtually disappears.

The observations stated in Lemma 2 lead to Propositions 2 and 3:

Proposition 2. *Consider the benchmark threshold model with no coordination issues among the citizens. The government's incentives to employ online commentators are non-monotonic with respect to the mean state of the world, $\bar{\theta}$:*

(i) For sufficiently low and sufficiently high values of the mean state of the world, online commentators are not employed by the government;

(ii) For moderate values of the mean state of the world, online commentators are employed only if μ is large enough, i.e. if the government cares sufficiently about the probability of regime change relative to the direct cost of employing online commentators.

Proof. The statement in part (i) for sufficiently low values of $\bar{\theta}$ follows from part (ii) of Lemma 2, while the statement for sufficiently high values of $\bar{\theta}$ follows from part (iv) of this lemma. Part (ii) is a consequence of part (ii) of Lemma 2 and the observations made in part (i) of the proposition. The weighting parameter, μ, must be high enough for the marginal benefit of increasing the payment, y, to outweigh the associated increase in the direct costs of employing the online commentators. □

For a better understanding of Proposition 2, recall that the higher is the online commentators' payment, y, the less informative is the private signal about the state of the world, θ, which makes the citizen put more weight on the mean state, $\bar{\theta}$, and less on the private signal, x, when forming her posterior. In the extreme, if the government increased the payment to infinity, the private signal would be completely uninformative for the citizen and the distribution of her posterior would coincide with the prior distribution of the state of the world, θ. If the prior mean state of the world is high enough relative to the critical state, θ^*, then by doing this, the government would ensure that that the citizen does not attack even for very low realisations of the private signal. This is because the citizen's posterior belief about the probability of the attack being successful would then be so low that condition (5) is not satisfied.

Conversely, if the prior mean state of the world is low relative to the critical state, θ^*, then the government would not want to make the citizen's private signal uninformative. Again, this would result in the citizen deciding whether to attack to regime based solely on the prior; however, in this case, if the prior mean state were low enough and the signal were imprecise, the citizen would decide to attack the regime regardless of her private signal. As a result, the government would be worse off. As long as the private information is precise enough, the bad prior mean may be outweighed by a good private signal, which may lead the citizen to refrain from attacking.

It follows that—in this benchmark threshold model with no coordination issues between the citizens—the larger is the mean state of the world, $\bar{\theta}$, relative to the exogenous critical state, θ^*, the more likely it is that increasing the payment above zero will make the government's benefit higher. Nevertheless, in the limit, as the mean state of the world becomes arbitrarily large, the marginal benefit from increasing the payment fades and virtually disappears. Since increasing the value of the payment is clearly associated with higher direct costs of employing the online commentators, the government will not choose to hire them if the mean state is sufficiently high. In fact, an increase in the mean state implies also that online commentators write more pro-government comments due to their intrinsic motivations, which pushes up the expected direct cost of employing online commentators.

Thus, for moderate—not too low and not too high—values of the mean state of the world, the government has the strongest incentives to employ online commentators. This will indeed happen if μ is large enough, i.e. if the government cares sufficiently about the probability of regime change relative to the direct cost of employing commentators.

Proposition 3. *Consider the benchmark threshold model with no coordination issues among the citizens. The government's incentives to employ online commentators are non-monotonic with respect to the ratio of the citizens' cost of attacking the regime to the portion of the payoff accessed only by attackers,* $\frac{c}{\delta}$*:*

(i) For sufficiently low and sufficiently high values of the ratio, $\frac{c}{\delta}$*, online commentators are not employed by the government;*

(ii) For moderate values of the ratio, $\frac{c}{\delta}$, online commentators are employed only if μ is large enough, i.e. if the government cares significantly more about the probability of regime change than about the direct cost of employing online commentators.

Proof. The reasoning here is similar to the one in the proof of Proposition 2. The statement in part (i) for sufficiently low values of $\frac{c}{\delta}$ follows from part (ii) of Lemma 2, while the statement for sufficiently high values of $\frac{c}{\delta}$ follows from part (iv) of this lemma. Part (ii) is a consequence of part (ii) of Lemma 2 and the observations made in part (i) of the proposition. As in Proposition 2, weighting parameter μ must be high enough for the marginal benefit of increasing the payment, y, to outweigh the associated increase in the direct costs of employing the online commentators. □

Note first that the citizen's cost of attacking, c, and the portion of the payoffs accessed only by the attackers, δ, appear in (6) only in the form of a ratio, $\frac{c}{\delta}$. Therefore, for the reason of brevity, we perform the analysis with respect to this ratio rather than the two parameters separately.

The reasoning here is similar to the one offered for Proposition 2. As the citizen's cost of attacking increases, setting the online commentator's payment at an arbitrarily high level becomes more desirable from the perspective of minimising the probability of regime change. This is because, if the cost of attacking is high enough, the government can ensure by doing this that condition (5) is not satisfied, i.e. that the posterior probability assigned by the citizen to the attack being successful is less than the critical value, $\frac{c}{\delta}$. Hence, regardless of her private signal, the citizen would not attack the regime. On the other hand, if the citizen's cost of attacking is relatively low, setting the payment at an arbitrarily high level would mean that the citizen attacks the regime for any private signal that she receives.

Thus, the government will choose to employ online commentators only if the citizen's cost of attacking is low enough. However, it must be noted that, as this cost becomes arbitrarily small, the marginal benefit from increasing their payment approaches zero. Consequently, the government will also not choose to hire them if the citizen's cost of attacking is sufficiently low.

It follows that the government's incentives to employ online commentators are strongest when the citizen's cost of attacking the regime takes moderate—not too low and not too high—values.

The purpose of this benchmark setting with no coordination issues between the citizens has been to provide better intuition for the transition between the linear model of Section 3 and the game of regime change analysed in Section 4. As it will become clear later, the mechanisms behind the results in Propositions 2 and 3 will remain essential in the general version of the threshold model.

4.3 Equilibrium

We now return to the general setup presented in Section 4.1, where the actions of the citizens are not perfectly coordinated.

In what follows, we limit our discussion to perfect Bayesian Nash equilibria in monotone strategies, that is, equilibria in which the citizens' strategies are non-decreasing in each citizen's private signal, x. The reason for this is twofold. First, the cumulative distribution function of a citizen's posterior about the state of the world, θ, is decreasing in her private signal, x. Furthermore, as long as the signal is sufficiently low (more precisely, if $x < \underline{x}$, where $Pr(\theta \leq 0 \mid \underline{x}) = \frac{c}{\delta}$), the citizens find it strictly dominant to attack. Conversely, if the signal is sufficiently high (more precisely, if $x > \bar{x}$, where $Pr(\theta \leq 0 \mid \bar{x}) = \frac{c}{\delta}$), it is strictly dominant for them not to attack.[21]

Definition 2. *An equilibrium in the threshold model consists of a critical state of the world and a critical private signal, (θ^*, x^*), such that a citizen attacks whenever her private signal is below the critical signal, $x \leq x^*$, and the regime is overthrown whenever the true state of the world is worse than the critical state, $\theta \leq \theta^*$.*

An alternative statement is that there is a critical state of the world, θ^*, 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 θ^*, resulting in a regime change. In other words, in this equilibrium, whenever $\theta < \theta^*$, at least θ players attack and the regime is overthrown. Conversely, whenever $\theta > \theta^*$, no more than θ players attack and the regime remains in place. Since the critical state of the world, θ^*, is determined endogenously in this version of the model, we will often denote it by $\theta^*(y; \cdot)$ to emphasise the fact that it is a function of the online commentators' payment, y, which is set by the government in the first stage of the game, as well as a number of exogenous parameters. It is also worth noting that nothing so far rules out multiple equilibria.

An equilibrium is determined by two conditions. The first equilibrium condition for determining the critical state, $\theta^*(y; \cdot)$, is sometimes referred to in the global games literature as the *payoff indifference condition* (e.g., Hellwig, 2002). In our model, a citizen is indifferent if the expected utility of attacking the regime is equal to the expected utility of refraining from an attack. In equilibrium, this happens when the citizen observes the critical private signal, x^*:

$$Pr\left(attack\, successful \mid x^*\right) - c = (1-\delta) Pr\left(attack\, successful \mid x^*\right)$$

$$\therefore Pr\left(attack\, successful \mid x^*\right) = \frac{c}{\delta}.$$

[21] Restricting attention to monotone Bayesian equilibria is common in the global games literature, for example, see Angeletos, Hellwig and Pavan (2007) and Loeper, Steiner and Stewart (2013).

With the structure of the equilibrium defined as above, the object of interest here is the probability that the true state of the world, θ, is smaller than or equal to the critical state, θ^*, given that the critical private signal, x^*, is observed by the citizen. Therefore, we have:

$$\begin{aligned} \frac{c}{\delta} &= Pr\left(attack\, successful \mid x^*\right) \\ &= Pr\left(\theta \leq \theta^* \mid x^*\right) \\ &= \Phi\left(\frac{\theta^* - E(\theta \mid x^*)}{\sqrt{Var(\theta \mid x^*)}}\right). \end{aligned} \tag{10}$$

Given the distribution of the state of the world, θ, conditional on a private signal, x, the citizen's indifference condition can be rewritten as:

$$\frac{c}{\delta} = \Phi\left(\sqrt{\frac{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}{(y^2\sigma_v^2 + \sigma_\varepsilon^2)\sigma_\theta^2}}\left(\theta^* - \bar{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\left(x^* - (\bar{\theta} + \bar{v}y)\right)\right)\right). \tag{11}$$

Thus, equivalently, the citizen is indifferent between attacking and not attacking if the private signal she observes is:

$$x^* = \frac{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}\theta^* - \frac{y^2\sigma_v^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}\bar{\theta} - \bar{v}y - \sqrt{\frac{(y^2\sigma_v^2 + \sigma_\varepsilon^2)(\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2)}{\sigma_\theta^2}}\Phi^{-1}(\frac{c}{\delta}). \tag{12}$$

The second equilibrium condition is often referred to in the global games literature as the *critical mass condition* (e.g., Hellwig, 2002). According to this condition, for a fundamental value of the state of the world, θ, that is equal to θ^*, the proportion of citizens who receive a signal weakly smaller than the critical value, x^*, should be equal in equilibrium to exactly θ^*. Mathematically, it means that:

$$\begin{aligned} \theta^* &= Pr\left(x \leq x^* \mid \theta^*\right) \\ &= \Phi\left(\frac{x^* - E(x \mid \theta^*)}{\sqrt{Var(x \mid \theta^*)}}\right) \\ &= \Phi\left(\frac{x^* - \theta^* - \bar{v}y}{\sqrt{y^2\sigma_v^2 + \sigma_\varepsilon^2}}\right). \end{aligned} \tag{13}$$

This can be rewritten as:

$$x^* = \theta^* + \bar{v}y + \sqrt{y^2\sigma_v^2 + \sigma_\varepsilon^2}\Phi^{-1}\left(\theta^*\right). \tag{14}$$

The two equilibrium conditions stated in equations (12) and (14) yield an equilibrium value of the critical state of the world, θ^*:

Proposition 4. *In the threshold model with coordination issues among the citizens, an equi-*

librium value of the critical state of the world, θ^*, solves:

$$\theta^*(y;\cdot) = \Phi\left(\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^*(y;\cdot)-\bar{\theta}\right) - \sqrt{\frac{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right). \tag{15}$$

4.4 Equilibrium Uniqueness

Since the setup of the model so far does not rule out multiplicity of equilibria, we now state a condition for equilibrium uniqueness, which makes further analysis more tractable.

Proposition 5. *The monotone equilibrium is unique if and only if $y \leq \sqrt{\frac{2\pi\left(\sigma_\theta^2\right)^2-\sigma_\varepsilon^2}{\sigma_v^2}}$.*

Proof. Substituting (14) into (12), we obtain the following condition:

$$\begin{aligned} U^{st}(\theta;\cdot) &\equiv 1-\Phi\left(\sqrt{\frac{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2(y^2\sigma_v^2+\sigma_\varepsilon^2)}}\left(\frac{y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}\left(\bar{\theta}-\theta\right)+\frac{\sigma_\theta^2\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}\Phi^{-1}(\theta)\right)\right)-\frac{c}{\delta} \\ &= 1-\Phi\left(\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}}\left(\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\bar{\theta}-\theta\right)+\Phi^{-1}(\theta)\right)\right)-\frac{c}{\delta}=0. \end{aligned} \tag{16}$$

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

$$\frac{\partial U^{st}(\theta;\cdot)}{\partial\theta} = -\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}}\phi(\cdot)\left(\frac{1}{\phi(\Phi^{-1}(\theta))}-\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\right) \tag{17}$$

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 θ is that $\sqrt{2\pi} \geq \frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}$, which can be rearranged to yield Proposition 5.[22] □

Proposition 5 is equivalent to the equilibrium uniqueness condition commonly seen in the global games literature: $\beta \geq \frac{\alpha^2}{2\pi}$, where $\beta = \left(y^2\sigma_v^2+\sigma_\varepsilon^2\right)^{-1}$ and $\alpha = \left(\sigma_\theta^2\right)^{-1}$ are the precisions of the private and the public (prior) signal, respectively.[23] In other words, there is a unique equilibrium value of the critical state of the world, θ^*, as long as the citizens' private signals are sufficiently precise. In our model, since this precision falls as the value of the online commentators' payment, y, increases, the payment must be low enough for the equilibrium to be unique. Moreover, if the noise variance σ_ε^2 is high enough, then there is no positive value of the payment that would satisfy the equilibrium uniqueness condition.

[22] We follow here the steps of the proof for the equilibrium uniqueness condition in Angeletos, Hellwig and Pavan (2007).

[23] See, for example, Angeletos, Hellwig and Pavan (2007).

For tractability reasons, we henceforth assume that the government wants to avoid at all costs the instabilities arising from multiple equilibria, and therefore it restricts its choice of the online commentators' payment to the interval $[0, y_H]$, where $y_H = \sqrt{\left(2\pi \left(\sigma_\theta^2\right)^2 - \sigma_\varepsilon^2\right) / \sigma_v^2}$. Nevertheless, because the government cares in our model also about the direct cost of employing the online commentators (which makes the optimal choice of the payment lower), this assumption will not determine the solution very often.

Crucially, Proposition 5 also turns out to be essential in determining the sign of the comparative statics effects, to the analysis of which we now proceed.

4.5 Comparative Statics of the Government's Benefit Function

In this section, we perform a comparative statics analysis of the government's benefit function with respect to (i) the mean state of the world, $\bar{\theta}$, and (ii) the ratio of the citizens' costs of attacking to the portion of payoffs accessed only by the attackers, $\frac{c}{\delta}$. We focus for now on the government's benefit function since the effects on the direct costs of employing online commentators are arguably rather straightforward. The analysis in this section is closely related to Metz (2002) and Heinemann and Bannier (2005), who have analysed the central bank's problem of determining rules for information dissemination in the context of currency crises.[24]

In the subsequent sections, we use the results of this section to analyse how changes in the above parameters affect the government's optimal choice of the online commentators' payment, and then we also provide a discussion of their impact on the direct cost of employing online commentators.

Lemma 3. *As the mean state of the world, $\bar{\theta}$, increases, it becomes less likely that the regime will be successfully overthrown, and thus the government's benefit becomes larger.*

Proof. The result is obtained by partially differentiating an equilibrium value of the critical state of the world, $\theta^* (y; \cdot)$, with respect to the mean state, $\bar{\theta}$, and then imposing the equilibrium uniqueness condition from Proposition 5. See the Appendix for the complete proof. □

The result in Lemma 3 is in line with common intuition. An increase in the mean state, $\bar{\theta}$, shifts the whole distribution of states to the right, which makes it less likely that the true state of the world is less than any given exogenously determined critical state, θ^*. Of course,

[24] Heinemann and Bannier (2005) analyse the central bank's optimal rules for transparency assuming that the transaction costs (equivalent to the citizens' costs of attacking in our model) are low compared with potential gains from a devaluation (equivalent to a regime change in our model). In contrast, we investigate the government's optimal choice of online commentators' payment (which is inversely related to transparency) for all possible levels of citizens' costs of attacking. Furthermore, unlike Heinemann and Bannier (2005), we also analyse how changes in the parameters of the model (more precisely, $\frac{c}{\delta}$ and $\bar{\theta}$) affect the government's optimal choice of the payment.

in this model, the critical state is determined endogenously, yet it turns out that—as long as the equilibrium uniqueness condition from Proposition 5 is satisfied—an increase in the mean state still makes it unambiguously less likely that the regime is overthrown. Since the strength of the regime is assumed here to be identical with the state of the world, an increase in the mean state could also be understood as an increase in the mean strength of the regime, which would provide an alternative interpretation for Lemma 3.

It is also worth noting that 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.

Lemma 4. *An increase in the ratio of the citizen's individual cost of attacking to the portion of the payoffs accessed only by the attackers, $\frac{c}{\delta}$, makes it less likely that the regime will be successfully overthrown, and thus it makes the government's benefit larger.*

Proof. The result is obtained by partially differentiating an equilibrium value of the critical state of the world, $\theta^*(y;\cdot)$, with respect to the ratio, $\frac{c}{\delta}$, and then imposing the equilibrium uniqueness condition from Proposition 5. See the Appendix for the complete proof. □

Similarly to Lemma 3, this result is in line with what common intuition would suggest. If there are large technical costs of participating in a protest, or participation is associated with a large opportunity cost of labour, then citizens will find it less worthwhile to participate in an attack against the current regime, thereby lowering the probability of a regime change. The intuition behind the effect of a decrease in δ—which in the model acts in the same way as an increase in c—is that the more citizens gain from a regime change even if they do not participate, the more tempting it is for them not to attack and hope that a sufficient number of others will attack. However, the more profitable this free-riding possibility is, the less likely it is that the current regime will be attacked and ultimately overthrown.

4.6 The Optimal Choice of Online Commentators' Payment

We now move to the heart of the analysis, that is, the government's incentives to employ online commentators and the optimal choice of the value of their payment.

Lemma 5. *Consider the threshold model with coordination issues between citizens. The government's marginal benefit from increasing the payment:*

(i) is equal to zero when the payment is zero and/or there is no uncertainty about the citizen's intrinsic valuation for money, σ_v^2;

(ii) is strictly positive if and only if $\theta^*(y;\cdot) < \bar{\theta} + \sqrt{\frac{\sigma_\theta^2(y^2\sigma_v^2+\sigma_\varepsilon^2)}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}}\Phi^{-1}(\frac{c}{\delta})$ and the conditions in part (i) do not hold, and it is negative otherwise;

(iii) approaches zero as the payment becomes arbitrarily large;

(iv) approaches zero for all values of y when at least one of $\bar{\theta}$ and $\frac{c}{\delta}$ becomes arbitrarily large.

Proof. By partially differentiating the expression for an equilibrium value of the critical state of the world, $\theta^*(y;\cdot)$, with respect to the payment, y, we obtain:

$$\frac{\partial\theta^*(y;\cdot)}{\partial y} = \frac{y\sigma_v^2\phi(\cdot)\left(\frac{1}{\sigma_\theta^2\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}\left(\theta^*-\bar{\theta}\right)-\frac{1}{\sqrt{\sigma_\theta^2\left(\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2\right)}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right)}{1-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}}, \tag{18}$$

where $\phi(\cdot)=\phi\left(\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^*-\bar{\theta}\right)-\sqrt{\frac{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right)$. Part (i) follows immediately.

For part (ii), 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 5 is satisfied, this ensures that the denominator is positive. However, the sign of the numerator is ambiguous; it is strictly positive if and only if

$$\frac{1}{\sigma_\theta^2\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}\left(\theta^*-\bar{\theta}\right)-\frac{1}{\sqrt{\sigma_\theta^2\left(\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2\right)}}\Phi^{-1}\left(\frac{c}{\delta}\right) > 0. \tag{19}$$

This can be rearranged to yield the result stated in part (ii).

Parts (iii) and (iv) follow from two properties of the normal distribution, $\phi'(x)=x\phi(x)$ and $\lim_{x\to-\infty}\phi'(x)=0$. We have that:

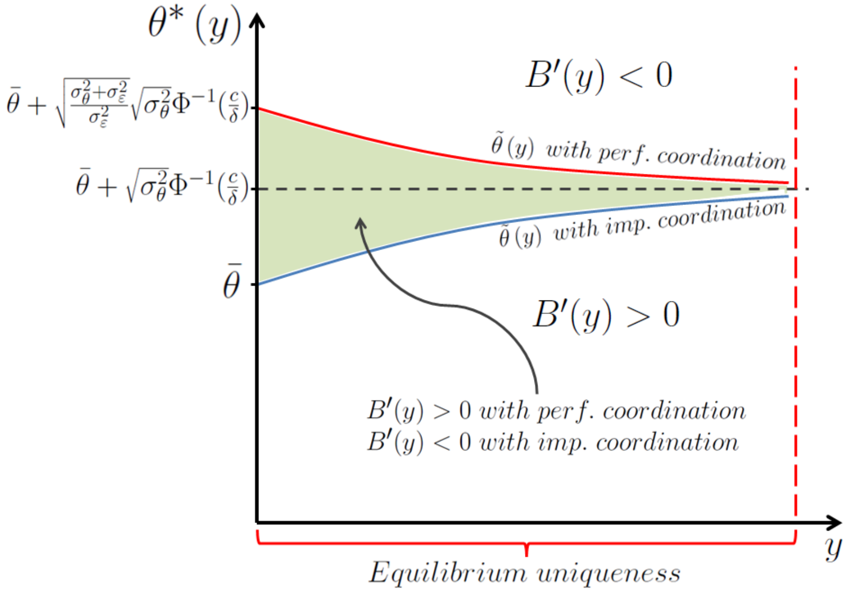
$$\lim_{\bar{\theta}\to\infty}\frac{\partial\theta^*(y;\cdot)}{\partial y}=\lim_{\bar{\theta}\to\infty}\frac{y\sigma_v^2\phi(\cdot)\left(\frac{1}{\sigma_\theta^2\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}\left(\theta^*(y;\cdot)-\bar{\theta}\right)-\frac{1}{\sqrt{\sigma_\theta^2\left(\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2\right)}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right)}{1-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}}=0,$$

where the last equality follows from two properties of the normal distribution, $\phi'(x)=x\phi(x)$ and $\lim_{x\to\pm\infty}\phi'(x)=0$. Since $B_P(y;\cdot):=\mu\Phi\left(\left(\bar{\theta}-\theta^*(y;\cdot)\right)\left(\sigma_\theta^2\right)^{-\frac{1}{2}}\right)$, this implies that as $\bar{\theta}$ becomes arbitrarily large, the government's marginal benefit from increasing y approaches zero for all values of y. The proof for $\frac{c}{\delta}$ follows the same steps. □

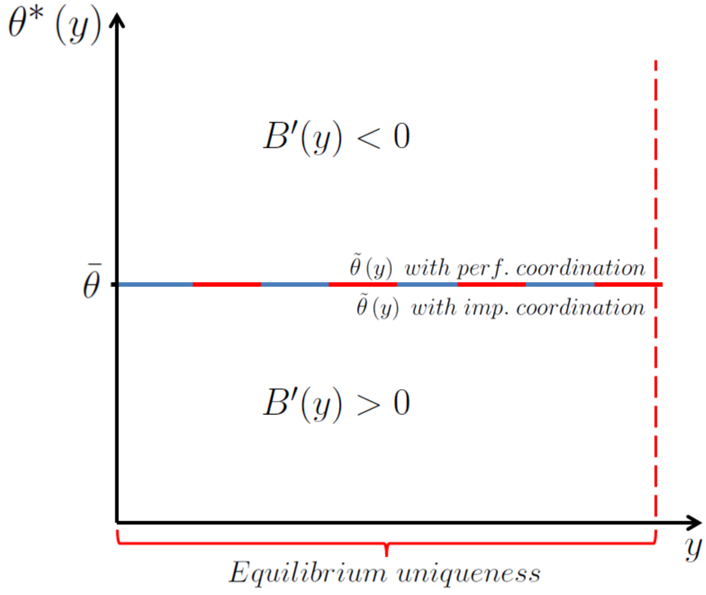
The basic intuition behind parts (i)-(iv) of Lemma 5 is the same as that behind parts (i)-(iv) of Lemma 2, which stated equivalent results in a benchmark setting with no coordination issues between citizens. The main implication remains that, in contrast to the linear model, the government's benefit may be increasing in the online commentators' payment, y. In the threshold model with coordination issues among the citizens, which is a richer setting than the perfect coordination benchmark, the threshold of positive marginal benefit is given by

$$\tilde{\theta}\left(y\right)=\bar{\theta}+\sqrt{\frac{\sigma_{\theta}^{2}\left(y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)}{\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}}\Phi^{-1}(\frac{c}{\delta}). \tag{20}$$

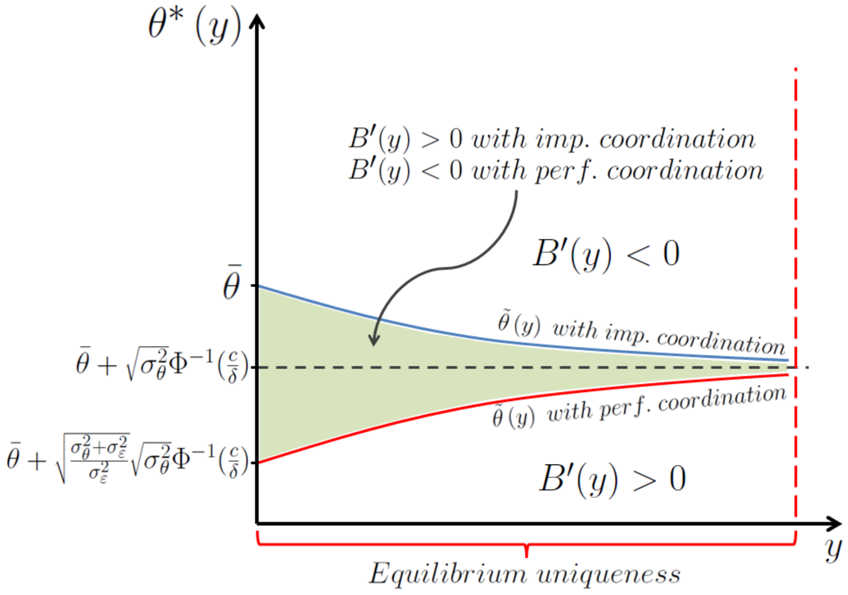
In Figure 1, we illustrate Lemma 5 by plotting the threshold of positive marginal benefit, $\tilde{\theta}\left(y\right)$, for the cases with perfect and imperfect coordination between citizens:


$\theta^{*}\left(y\right)$
$\bar{\theta}+\sqrt{\frac{\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2}}{\sigma_{\varepsilon}^{2}}}\sqrt{\sigma_{\theta}^{2}}\Phi^{-1}(\frac{c}{\delta})$
$\bar{\theta}+\sqrt{\sigma_{\theta}^{2}}\Phi^{-1}(\frac{c}{\delta})$
$\bar{\theta}$
$B'(y)<0$
$\tilde{\theta}\left(y\right)$ with perf. coordination
$\tilde{\theta}\left(y\right)$ with imp. coordination
$B'(y)>0$
$B'(y)>0$ with perf. coordination
$B'(y)<0$ with imp. coordination
y
Equilibrium uniqueness

Case 1: Costs are relatively high: $\frac{1}{2}\delta<c<\delta$ (which implies $\Phi^{-1}(\frac{c}{\delta})>0$).


$\theta^{*}\left(y\right)$
$B'(y)<0$
$\bar{\theta}$
$\tilde{\theta}\left(y\right)$ with perf. coordination
$\tilde{\theta}\left(y\right)$ with imp. coordination
$B'(y)>0$
y
Equilibrium uniqueness

Case 2: Costs are intermediate: $c = \frac{1}{2}\delta$ (which is equivalent to $\Phi^{-1}(\frac{c}{\delta}) = 0$)



Case 3: Costs are relatively low: $c < \frac{1}{2}\delta$ (which is equivalent to $\Phi^{-1}(\frac{c}{\delta}) < 0$).

Figure 1. The sign of the government's marginal benefit from increasing the online commentators' payment as a function of the critical state, θ^*, and the online commentators' payment, y, for different values of the citizen's cost of attacking, c: a comparison of the general model with the benchmark threshold model with no coordination issues between citizens.

It follows from Figure 1 that, when the citizen's costs of attacking are relatively high, $\frac{1}{2}\delta < c < \delta$, the government's marginal benefit from increasing online commentators' payment is more likely to be positive in the benchmark threshold model with no coordination issues among the citizens. More precisely, there is a parameter space in the critical state of the world, θ^*, and the payment, y, where the government's marginal benefit from increasing the payment is negative in the general model and positive in the benchmark model with no coordination issues among the citizens. On the other hand, when the costs of attacking are relatively low, $c < \frac{1}{2}\delta$, the reverse is true: the government's marginal benefit from increasing the payment is more likely to be positive when the citizens are not perfectly coordinated. Finally, if the costs of attacking are intermediate, $c = \frac{1}{2}\delta$, the respective parameter spaces in θ^* and y are exactly the same in the benchmark and the general model.

It is worth noting that, at a broad level, the results in the general and the benchmark threshold model are similar. Although the sign of the government's marginal benefit of the payment does change for some parameter spaces when coordination issues between citizens are introduced, the main implication—that it may be optimal for the government to employ online commentators—remains unchanged. In other words, the main mechanisms present in the benchmark threshold model carry over to the general model discussed in this section.

Nevertheless, the comparative statics result of Lemma 5 is only a first step towards finding the value of the online commentators' payment that would maximise the government's utility. This is because, in this threshold model, both the critical state of the world, $\theta^* (y; \cdot)$, and the threshold of positive marginal benefit, $\tilde{\theta}(y)$, are functions of the online commentators' payment, y. The following definition will prove useful in the subsequent discussion:

Definition 3. *The intrinsic fragility of the regime, $\theta^*(0; \cdot)$, is the critical state of the world that arises when online commentators are not employed by the government, i.e. when their payment is zero.*

We proceed with the analysis as follows: we first look at the intrinsic fragility of the regime, $\theta^*(0; \cdot)$, and then we use the comparative statics results of Lemma 5 to determine how the government's benefit is affected as the payment to online commentators is increased. This analysis yields the following result:

Lemma 6. *There exist certain thresholds of intrinsic fragility of the regime τ and τ', with $\tau \leq \tau'$, such that:*

(i) If the intrinsic fragility of the regime is greater than τ', the lowest possible payment to online commentators, $y = 0$, maximises the government's benefit;

(ii) If the intrinsic fragility of the regime decreases below τ, the payment which maximises the government's benefit increases discontinuously and takes the highest possible value that is permitted by the equilibrium uniqueness condition, $y = y_H$.

Proof. For a general idea of the proof, note that if the intrinsic fragility of the regime, $\theta^* (0; \cdot)$, is high relative to the mean state of the world, then the probability of regime change is increasing in the online commentators' payment, y, for all values of y. Equivalently, this means that the government's benefit is unequivocally decreasing in y. On the other hand, if the intrinsic fragility of the regime is low relative to the mean state of the world, the probability of regime change is unequivocally decreasing in y. For the complete proof with an analysis of the discontinuities, see the Appendix. □

The result stated in Lemma 6, however, seems counterintuitive as one would think that regimes that are intrinsically fragile would have stronger incentives to manipulate information, rather than the other way round. The intuition behind this proves to be closely related to the reasoning behind part (ii) of Lemma 2, which leads to qualitatively similar results in a benchmark threshold model with no coordination issues between the citizens.

When the intrinsic fragility, $\theta^* (0; \cdot)$, is low relative to the prior mean, $\bar{\theta}$, the citizens' prior incentives to attack are weak. Since the fundamentals follow a normal distribution, there is always a positive probability that the state of the world turns out to be lower than the critical state of the world, $\theta^* (0; \cdot)$. However, once the citizens' private signals are made less precise, the citizens attach less weight to them, and thus their posteriors are centred more closely around the prior mean, which is here assumed to be high relative to the measure of intrinsic fragility. This leads the citizens to refrain from attacking even for very bad states, which in turn results in a lower probability that a successful regime change will take place. Therefore, since an increase in the online commentators' payment makes the private signals less precise in this model, the government's incentives to offer a strictly positive payment to online commeantators are higher when the intrinsic fragility, $\theta^* (0; \cdot)$, is low.

On the other hand, if the intrinsic fragility of the regime, $\theta^* (0; \cdot)$, is high relative to the prior mean, $\bar{\theta}$, the citizens' prior incentive to attack is rather strong and the argument is converse. When the citizens receive precise private signals, they are more likely to realise that the true state of the world, θ, is greater than the critical state, $\theta^* (0; \cdot)$. As a result, they are less likely to attack the regime, which clearly increases the probability that the regime will survive. Therefore, in this case, making the online commentators' payment as low as possible is preferable from the government's point of view.

It is also worth noting that the government's incentives to employ online commentators are characterised by discontinuities. The nature of these discontinuities differs somewhat depending on the citizens' relative costs of attacking the regime. When the costs are intermediate or relatively high, $\frac{1}{2}\delta \leq c < \delta$, then the thresholds τ and τ' coincide, $\tau = \tau'$. Consequently, the choice of the payment that maximises the government's benefit is characterised by a simple cut-off result. As the regime's intrinsic fragility reaches a certain low level, the payment that maximises the government's benefit increases discontinuously from $y = 0$ to $y = y_H$. The picture is slightly different when the citizens' costs of attacking the regime are relatively low, $c < \frac{1}{2}\delta$. In that case, whenever the intrinsic fragility of the regime is between τ and τ', the government's benefit may be maximised for a value of the payment that is strictly within the interval $(0, y_H)$. As a result, once a certain low level of the regime's intrinsic fragility is reached, the payment that maximises the government's benefit increases discontinuously from $y \in [0, y_H)$ to $y = y_H$. The main implication remains here, however, that the government's incentives to employ online commentators may increase dramatically once the regime's intrinsic fragility becomes low enough.[25]

Proposition 6. *The government's incentives to employ online commentators are non-monotonic with respect to the mean state of the world, $\bar{\theta}$:*

(i) For sufficiently low and sufficiently high values of the mean state of the world, online commentators are not employed by the government;

(ii) For moderate values of the mean state of the world, online commentators are employed

[25] A more detailed discussion of the discontinuities is provided in the proof of Lemma 6 in the Appendix.

only if μ is large enough, i.e. if the government cares sufficiently about the probability of regime change relative to the direct cost of employing online commentators.

Proof. The statement in part (i) for sufficiently low values of $\bar{\theta}$ follows from Lemma 3 and part (ii) of Lemma 6, whereas the statement on sufficiently high values of $\bar{\theta}$ follows from part (iv) of Lemma 5. Part (ii) of Proposition 6 is a consequence of part (ii) of Lemma 6 and the statement in part (i) of the proposition. □

The statement in Proposition 6 is parallel to Proposition 2, which states an equivalent result in the benchmark threshold model with no coordination issues between the citizens. As before, we could interpret an increase in the mean state of the world, $\bar{\theta}$, as an improvement in the general state of the country. Note here that, by Lemma 3, the intrinsic fragility of the regime, $\theta^* (0; \cdot)$, is strictly decreasing in the mean state, $\bar{\theta}$. An increase in the mean state implies also that the threshold of positive marginal benefit, $\tilde{\theta}(y)$, shifts upwards. Thus, we obtain the seemingly counterintuitive result that the higher is the mean state of the world, the more likely it is that setting a strictly positive value of the online commentators' payment will maximise the government's benefit.

However, it must also be noted that as the mean state of the world, $\bar{\theta}$, becomes arbitrarily large, the government's marginal benefit from increasing the online commentators' payment approaches zero for all values of the payment, y. This is because, when the mean state is very high, then by Lemma 3 the probability that the regime is successfully overthrown is very low, and therefore an additional increase in the payment brings little benefit for the government. At the same time, an increase in the payment clearly leads to higher costs of employing the online commentators. In fact, an increase in the mean state makes these costs even higher since then, on average, online commentators write more pro-government comments due to the intrinsic motivations, and they have to be paid for them, too.

Thus, we obtain a result that for moderate—not too high and not too low—values of the mean state of the world, the government may find it optimal to employ online commentators. The government will indeed choose to hire them if μ is large enough, i.e. if it cares sufficiently about the probability of regime change relative to the direct cost of employing the commentators.

Proposition 7. *The government's incentives to employ online commentators are non-monotonic with respect to the ratio of the citizens' cost of attacking the regime to the portion of the payoff accessed only by attackers, $\frac{c}{\delta}$:*

(i) For sufficiently low and sufficiently high values of the ratio, online commentators are not employed by the government;

(ii) For moderate values of the ratio, online commentators are employed only if μ is large enough, i.e. if the government cares significantly more about the probability of regime change than about the direct cost of employing online commentators.

Proof. The reasoning here is very similar to that in the proof of Proposition 6. The statement in part (i) for sufficiently low values of $\frac{c}{\delta}$ follows from Lemma 4 and part (ii) of Lemma 6, whereas the statement on sufficiently high values of $\frac{c}{\delta}$ follows from part (iv) of Lemma 5. Part (ii) of Proposition 7 is a consequence of part (ii) of Lemma 6 and the statement in part (i) of the proposition. □

This result is parallel to Proposition 3, which makes an equivalent statement for the benchmark threshold model with no coordination issues between the citizens. By Lemma 4, the intrinsic fragility of the regime, $\theta^* (0; \cdot)$, is strictly decreasing in ratio $\frac{c}{\delta}$. Furthermore, a rise in $\frac{c}{\delta}$ means also that the threshold of positive marginal benefit, $\tilde{\theta} (y)$, shifts upwards. By the arguments of Lemma 6, it follows that the higher is the citizens' cost of attacking relative to the portion of payoffs accessed only by attackers, the more likely it is that setting a strictly positive value of the online commentators' payment will maximise the government's benefit.

Even though the critical state of the world, θ^*, is now endogenously determined, the main mechanism behind the impact of the citizens' cost of attacking on the optimal choice of the online commentators' payment carries over from the benchmark model with no coordination issues among the citizens. The higher is the cost of attacking, the more likely it is—other things being equal—that the citizen does not attack with only prior information. Nevertheless, with precise private information, a citizen could receive a low private signal that would induce her to attack the regime even though she would not do it based only on the prior. Thus, employing online commentators, and thereby making private information uninformative, is especially effective when the cost of attacking is high. On the other hand, when the cost of attacking is so low that the citizens would attack the regime with only prior information, the government may prefer to make the private signals precise. This is because, given precise private information, a citizen may refrain from attacking if she receives a sufficiently high signal about the state of the world—whereas she would not abstain from attacking based only on her prior.

Nevertheless, as the citizen's costs of attacking the regime become arbitrarily large, the government's marginal benefit from increasing the online commentators' payment approaches zero for all values of the payment, y. The intution here is very similar to the one provided for Proposition 6. When the costs of attacking are very high, then by Lemma 4 the probability that the regime is successfully overthrown is very low, and an additional increase in the payment brings little benefit for the government. At the same time, an increase in the payment clearly leads to higher costs of employing the online commentators.

Therefore, the government is most likely to employ online commentators when the ratio of the citizens' cost of attacking the regime to the portion of the payoff accessed only by attackers, $\frac{c}{\delta}$, takes moderate values. This will indeed happen if μ is large enough, i.e. if the government cares sufficiently about the probability of regime change relative to the direct cost of employing the commentators.

4.7 Discussion

The results of the linear model, which was discussed in Section 3, and of the threshold model presented here can be related to the main empirical result of King et al. (2013), which states that, in the context of China's censorship programme, the theory of collective action potential is correct while the state critique theory is not. In other words, the censoring activities of the Chinese government are not aimed at restricting any criticism of the government and its policies, but instead their purpose is to prevent collective action.

According to the linear model—which is our economic interpretation of the state critique theory—the optimal payment is always equal to zero, which means that the government has no incentives to employ online commentators. On the other hand, in the threshold model, where the government's payoff depends on the probability that a sufficient proportion of citizens attack the regime, the government's marginal benefit from increasing the payment—and thus the optimal choice of the payment—may be strictly positive. Hence, according to the model, if the government employs online commentators, it does so because it cares about the probability of a collective action rather than about the level of criticism *per se*.

The model thus provides *a* theoretical rationale for the empirical finding of King et al. This implication rests nonetheless on one crucial assumption, which is that the citizens observe the value of online commentators' payment. More generally, this means that, in the model presented so far, the citizens have been assumed to be fairly well-informed about the regime's manipulative practices. Therefore, a more precise implication of the model is that fear of state critique should not be a reason for the government to employ online commentators so long as the citizens are sophisticated enough to know the value of the online commentators' payment.

5 What If Online Commentators' Payment Is Not Observed by Citizens?

In this section, we relax—for the linear model—the assumption that the citizens observe how much the online commentators are being paid for every pro-government comment. In the real world, the Chinese citizens are now becoming increasingly aware of the existence of online commenatotors and, in fact, the term "fifty center" has become a common derogatory expression for anyone who actively and publicly posts opinions online that defend or support government policy.[26,27] However, when the system of paid online commentators was intro-

[26] http://chinadigitaltimes.net/space/Fifty_cents

[27] http://chinadigitaltimes.net/2010/08/an-inside-look-at-a-50-cent-party-meeting/

duced, few people knew about it. In this section, we investigate the government's incentives to employ online commentators when the citizens are not so well-informed as to know the value of the commentators' payment.

This version of the linear model shares many similarities with models of managerial career concerns, e.g. Holmström (1999), and Dewatripont, Jewitt and Tirole (1999a, 1999b). The Appendix provides a more detailed discussion of this connection. The main result is as follows:

Proposition 8. *Suppose that the utility of the government is linear in the citizen's posterior expectation and the value of the online commentators' payment, y, is not observed by the citizens. Then the optimal payment, y^*, is strictly positive if and only if $\mu\bar{v}\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2} > \bar{\theta}$. If this condition holds, y^* solves the condition $\frac{\mu\sigma_\theta^2\bar{v}}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2} = \bar{\theta} + 2\bar{v}y$.*

Proof. Given her private signal, x, and a conjecture about the online commentators' payment, $\hat{y}$, the citizen's expectation of the state of the world, θ, is:

$$E\left(\theta \mid x, \hat{y}\right) \quad := \quad \overline{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \hat{y}^2\sigma_v^2 + \sigma_\varepsilon^2}\Big(x - (\overline{\theta} + \bar{v}\hat{y})\Big) \text{ where } x := a + \varepsilon. \tag{21}$$

The government's utility function is then:

$$\begin{aligned} U_P(y, \hat{y}; \cdot) &= E_{\theta,v,\varepsilon}\left[\mu E\left(\theta \mid x, \hat{y}\right) - ay\right] \\ &= E_{\theta,v,\varepsilon}\left[\mu\left(\overline{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \hat{y}^2\sigma_v^2 + \sigma_\varepsilon^2}\Big(x - (\overline{\theta} + \bar{v}\hat{y})\Big)\right) - ay\right] \\ &= \mu\left(\overline{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \hat{y}^2\sigma_v^2 + \sigma_\varepsilon^2}\left(\bar{v}(y - \hat{y})\right)\right) - (\bar{\theta} + \bar{v}y)y. \end{aligned} \tag{22}$$

The government chooses the value of the online commentators' payment, y, so as to maximise its utility given the conjecture, $\hat{y}$. The first order condition with respect to y satisfies

$$\frac{\mu\sigma_\theta^2\bar{v}}{\sigma_\theta^2 + \hat{y}^2\sigma_v^2 + \sigma_\varepsilon^2} = \bar{\theta} + 2\bar{v}y, \tag{23}$$

where the left-hand side of the equation can be interpreted as the government's marginal benefit of increasing y (given conjecture $\hat{y}$), while the right-hand side can be seen as the corresponding marginal cost. In equilibrium, the citizen's conjecture is true, which means that $\hat{y} = y$. This yields the statement in the proposition. □

The most important implication of Proposition 8 is that, when the online commentators' payment, y, is not observed by the citizens, the government may have a signal-jamming incentive to set a strictly positive value of y even if its benefit function is linear in the citizens' posteriors. This contrasts with the result of the linear model presented in Section 3, where the value of the payment was observed by the citizens. Thus, one potentially crucial factor in the empirical results of King et al. (2013) is how well-informed the citizens are about the regime's manipulative practices.

Moreover, Proposition 8 implies that, for the optimal value of the payment, y, to be strictly greater than zero, the following must hold: (i) μ, the relative importance of the benefit component in the government's utility function, must be high enough; (ii) $\bar{v}$, the commentator's average intrinsic valuation for money, must be high enough; (iii) σ_ε^2, the variance of the noise in the citizen's private signal, must be low enough; (iv) σ_θ^2, the variance of the state of the world, must be low enough; and (v) $\bar{\theta}$, the average state of the world, must be low enough.

Furthermore, we can say that the optimal value of the payment, y^*, is unique and stable. The uniqueness result follows from the fact that the marginal benefit is decreasing in y whereas the marginal cost is increasing, so if the marginal benefit and the marginal cost curves cross, they do it only once. Clearly, if they do cross, the marginal benefit curve is above (below) the marginal cost curve for $y < (>) y^*$, which means that y^* is always stable.

The following figure illustrates the government's marginal benefit and marginal cost functions of increasing the online commentators' payment:

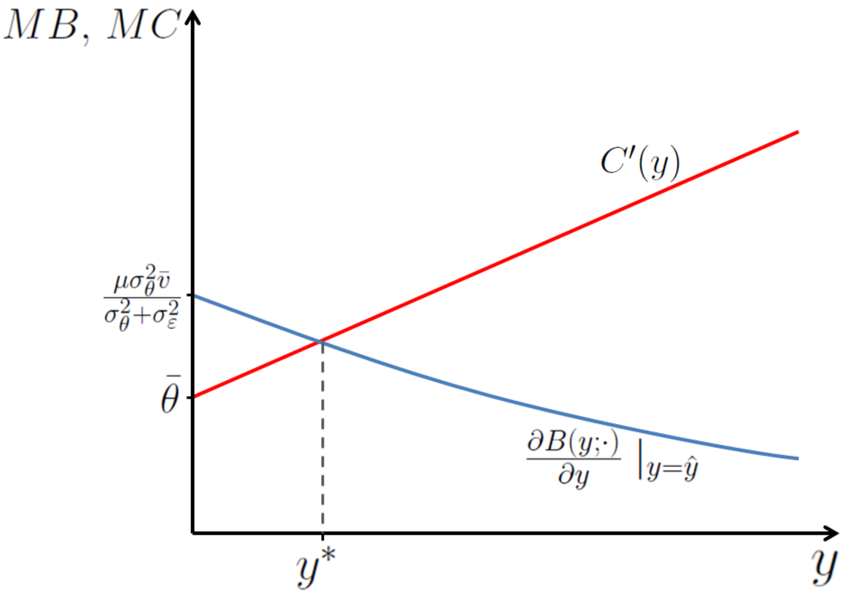


Figure 2. The equilibrium choice of the online commentators' payment in the linear model when its value is not observed by the citizens.

Interestingly, according to the linear model in which the payment is not observed by the citizens, an increase in the mean state of the world, $\bar{\theta}$, makes it less likely that the optimal value of the online commentators' payment will be strictly positive. This result is different from the one we had in the threshold model that was discussed in Section 4, where an increase in $\bar{\theta}$

could make it more likely that the government will employ online commentators. This would occur in the threshold model if μ were high enough, i.e. if the government cared significantly more about the probability of regime change than about the direct cost of employing the commentators. The result in this version of the linear model follows from the fact that an increase in the mean state of the world shifts the marginal cost curve upwards (more posts are written by commentators for non-pecuniary reasons), whereas the effect on the marginal benefit is fully internalised by rational citizens who are informed about $\bar{\theta}$.

We now make a number of further observations about the model presented in this section:

Corollary 1. *If the citizens' mean intrinsic valuation for money is zero, $\bar{v} = 0$, then the government wishes to set the online commentators' payment, y, at zero even if y is not observed by the citizen.*

Thus, when the commentators' mean intrinsic valuation for money is zero, the government's incentives under observability and non-observability of the payment coincide. This follows from the fact that, when $\bar{v} = 0$, online commentators—on average—do not respond to an increase of the payment by writing more pro-government comments. As a result, the government has no signal-jamming incentive to increase the value of the payment.

Corollary 2. *If the variance of the intrinsic valuation for money is zero, i.e. $\sigma_v^2 = 0$, then the marginal benefit of increasing y is independent of y. The optimal choice of the online commentators' payment is then $y^* = \frac{1}{2}\left(\mu \frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2} - \frac{\bar{\theta}}{\bar{v}}\right)$.*

In other words, when there is no uncertainty about the commentators' intrinsic valuation for money, we obtain a result that is reminiscent of the pure additive model of career concerns in Dewatripont et al. (1999b), where the signal-to-noise ratio is independent of effort.

In addition to this, we can also perform a comparative statics analysis to see how the value of the online commentators' payment is affected by different features of the environment:

Corollary 3. *Suppose that the utility of the government is linear in the citizen's posterior expectation and the online commentators' payment is not observed by citizens. The optimal value of the payment, y^*, rises as:*

(a) the government cares more about the benefit relative to the direct cost, i.e. parameter μ increases;

(b) the mean state of the world, $\bar{\theta}$, falls;

(c) the variance of the state of the world, σ_θ^2, increases;

(d) the variance of the intrinsic valuation for money, σ_v^2, falls;

(e) the variance of the noise in the citizen's private signal, σ_ε^2, falls;

(f) as the commentators' average intrinsic valuation for money, $\bar{v}$, increases.

The intuition behind (a) is that an increase in μ raises the marginal benefit of increasing the payment, but at the same time it does not affect the marginal cost.

Part (b) follows from the fact that an increase in $\bar{\theta}$ shifts the marginal cost curve upwards (as more posts are written by commentators for non-pecuniary reasons) whereas the effect on the marginal benefit is fully internalised by rational citizens who are informed about the mean state of the world.

The result in (c) arises because an increase in σ_θ^2 makes $Cov(\theta, x)$ higher for any given level of the online commentators' payment, y, and since the marginal benefit from increasing y is proportional to $Cov(\theta, x)$, it also becomes higher for any given level of y.

As far as (d) is concerned, an increase in σ_v^2 leads $Cov(\theta, x)$ to decrease at a faster rate as the payment is increased, and since the marginal benefit of increasing y is proportional to $Cov(\theta, x)$, it decreases at a faster rate as well.

An increase in σ_ε^2 means that there is an increase in the noise in the citizen's private signal that is not controlled by the government. Since this noise leads to a fall in the marginal benefit of increasing the payment, the result in (e) follows.

The impact of an increase in $\bar{v}$ on the optimal choice of the online commentators' payment is more complex since, on the one hand, it raises the marginal benefit of increasing the payment, but at the same time it makes the marginal cost increase at a faster rate. The result in part (f) is shown mathematically in Lemma 8.

Lemma 8. *Let $\xi = \frac{\sigma_\varepsilon^2}{\sigma_\theta^2}$ and $\eta = \frac{\sigma_v^2}{\sigma_\theta^2}$. As $\bar{v}$ increases from $\bar{\theta}(1+\xi)$ to infinity, the equilibrium value of the payment y starts from zero and approaches $\bar{y}$, where $\bar{y}$ is the (real) root of the cubic equation $y^3 + \frac{1+\xi}{\eta} y - \frac{\mu}{2\eta} = 0$.*

Proof. See the Appendix. □

Finally, it is worth noting that the government is actually worse off when the online commentators' payment, y, is not observed by the citizens. If the payment were observed, the government would optimally choose $y^* = 0$ and its ex ante utility would be equal to $\mu\bar{\theta}$. On the other hand, unobservability of the payment may drive the government to increase it above zero, for any conjecture $\hat{y}$ held by the citizens. In equilibrium, the ex ante benefit of the government remains nonetheless unaffected, which is a result of the benefit function being linear in the citizen's posterior. This ex ante benefit is $\mu\bar{\theta}$, that is, it is the same as in the case where the payment is observable. Since the cost clearly increases as the payment is raised above zero, the government is worse off whenever the optimal signal-jamming incentive is so strong that the government is induced to set a strictly positive value of the online commentators' payment. Thus, the government would be better off if it only could commit to a payment of zero.[28]

[28] This result is also present in signal-jamming models of managerial career concerns.

6 Conclusion

With the rapid growth of social media and the increasingly ubiquitous use of smartphones, news spread widely and quickly enough that those authoritarian regimes which desire to manipulate information face the risk of losing credibility if they simply censor the uncomfortable pieces. Consequently, more subtle ways of manipulating information are needed. One possible solution for those regimes is to employ online commentators whose task is to post pro-government opinions on the Internet. In fact, as the evidence provided by the Freedom House shows, this has become a common method of influencing the information content available to citizens.

We have formally investigated the regimes' incentives to hire such agents in a model with three types of players: the government, online commentators, and citizens. In the model, the government offers to pay a certain wage to the online commentators for every pro-regime opinion that they post on the Internet. In addition to this extrinsic motivation, the commentators have also intrinsic motives to write positively about the government if the social and economic situation in the country is genuinely good. Importantly, at the moment of contracting, the government is uninformed about the precise intrinsic and extrinsic motivations of any particular commentator. Citizens observe the number of posts written by the commentators and, on this basis, they update their beliefs about the state of the country.

Within the model, we have constructed two distinct settings, which are designed to correspond to two theories—proposed by King et al. (2013)—of what constitutes the goals of the Chinese regime as implemented in their censorship programme: the state critique theory and the theory of collective action potential. By doing this, we have extrapolated the empirical findings of King et al. to the programme of employing online commentators. In both settings, the government's utility function consists of a benefit function less the direct cost of employing the online commentators, with a parameter introduced to indicate the relative importance of the benefit component in the utility function. However, the two environments differ with respect to the actions taken by the citizens based on their posteriors, and with respect to the benefit functions of the government.

In the first setting, the government's benefit function has been linear in the citizens' posteriors about the state of the world. In this environment, which is our economic interpretation of the state critique theory of King et al., we have shown that the government has no incentives to employ online commentators as long as the citizens are sophisticated enough to know the value of the online commentators' payment.

In the second setting, we have assumed that the citizens decide—based on their posteriors—whether to attack the regime or not, and that the government's benefit function is the ex ante probability that the collective attack is unsuccessful. This environment aims to illustrate the theory of collective action potential, as defined by King et al. We have demonstrated that, in contrast to the linear setting, the government may now find it optimal to employ online commentators in order to reduce the likelihood of an attack on the regime (even if the citizens are so well-informed about the regime's manipulative practices as to know the value of the commentators' payment). Thus, one contribution of our paper has been to provide a theoretical rationale for the empirical result of King et al. that the theory of collective action potential is correct while the state critique theory is not.

Furthermore, we have shown in the collective action environment that the incentives of regimes to employ online commentators are non-monotonic in (i) the citizens' private costs of attacking the regime, (ii) the portions of the gains from regime change that are accessed only by those who had attacked the initial regime, and (iii) the prior mean state of the world, which could be interpreted as the general state of the country. The regimes are most likely to employ online commentators when these parameters take moderate values. Interestingly, these results suggest that, in fact, we may observe the system of online commentators become more widespread as the social and economic situation improves or as the costs of protesting increase.

Finally, we have investigated the possibility that the citizens are not sophisticated enough to know the value of the online commentators' payment. We have demonstrated that the government may then find it optimal—due to a signal-jamming incentive—to employ the commentators even if it cares about state critique rather than the possibility of a collective action. Thus, one could argue that fear of state criticism could have been a valid reason to introduce the system of online commentators in China more than a decade ago, when the citizens were less informed about the regime's manipulative practices.

Obviously, the model presented in this paper has certain limitations. First, it must be noted that employing online commentators is one method of manipulating information, but there are many more. For example, as the Freedom House lists, governments may also (i) block social media platforms or communication applications, (ii) block or censor political, social and religious websites, (iii) attack or arrest online journalists and bloggers for political writings, or even (iv) shut down local or nationwide networks.[29] In the economics and formal political science literature, Lorentzen (2014) and Shadmehr and Bernhardt (2015) analyse the incentives of regimes to censor information available to citizens, while government control of media is the subject of, e.g., Besley and Prat (2006), Egorov, Guriev and Sonin (2009), and Gehlbach and Sonin (2014).

In our model, the government's incentives to employ online commentators are analysed in isolation from other methods of counteracting regime change and, as a result, any potential interactions with them are disregarded. A particularly interesting extension would be to

[29] Freedom House, 2013, "Freedom on the Net 2013: A Global Assessment of Internet and Digital Media" (2013), edited by S. Kelly, M. Truong, M. Earp, L. Reed, A. Shahbaz, and A. Greco-Stoner; https://freedomhouse.org/report-types/freedom-net

investigate to what extent the government's incentives to employ online commentators change when also censorship—a more orthodox technique of information manipulation—is available. In particular, one could analyse, along the lines of Holmström and Milgrom's (1994) analysis in the context of firms, whether these methods are complementary instruments for reducing the probability of a regime change.[30]

Second, we have assumed in the model that the government's choice of the payment does not affect the pool of online commentators who are willing to work. More realistically, one would expect to see some degree of self-selection among online commentators. For instance, it could be argued that those who accept the government's offer are likely to have, on average, higher intrinsic valuations for money and lower costs of writing posts when compared to a typical citizen.

Furthermore, one could investigate how the robustness of the model would be affected if we relaxed the assumption of a one-to-one relationship between citizens and commentators. Since the number of online commentators is in reality lower than the number of citizens by at least an order of magnitude, it would be interesting to see whether the predictions of the model change if there is only a finite number of commentators. Consequently, they would be listened by more than just one citizen, which would make the correlation between citizens' private signals strictly positive. We conjecture that this could aggravate the possibility of multiple equilibria.

In our model, we have analysed how the government's incentives are affected in the linear model when the citizens do not observe the value of the online commentators' payment. An interesting avenue for research would be to investigate also how the government's incentives change in the global games environment of the threshold model.

Lastly, one could also compare the regimes' incentives to employ online commentators with the incentives of firms which hire paid reviewers. This would allow us to see whether there are any important similarities or differences between the two phenomena, which could help further elucidate the incentives of governments to manipulate information.

[30] The interplay between a number of methods of preventing a regime change—but not including the possibility of employing online commentators—is analysed by Guriev and Treisman (2015). In their model, a dictator can invest in making convincing state propaganda, censoring independent media, co-opting the elite, or equipping police to repress attempted uprisings. They show that censorship and co-optation of the elite are substitutes, but both are complements of propaganda, whereas repression of protests is a substitute for all the other techniques.

7 Appendix

7.1 Omitted Proofs

Lemma 1

Proof. Noting that the variance of θ conditional on x is

$$Var(\theta \mid x) = \frac{(y^2\sigma_v^2 + \sigma_\varepsilon^2)\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2},$$

we can express the probability that the attack is successful as follows:

$$\begin{aligned} Prob(attack\, successful \mid x) &= \Phi\left(\frac{\theta^* - E(\theta \mid x)}{\sqrt{Var(\theta \mid x)}}\right) \\ &= \Phi\left(\frac{\theta^* - \overline{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\left(x - (\overline{\theta} + \bar{v}y)\right)}{\sqrt{\frac{(y^2\sigma_v^2 + \sigma_\varepsilon^2)\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}}}\right), \end{aligned} \tag{24}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for the normal distribution.

Now, let x^* be a cut-off value of the citizen's private signal such that the citizen attacks the regime whenever $x \leq x^*$ and refrains from attacking whenever $x > x^*$. In other words, the critical private signal, x^*, is determined by the citizen's indifference condition:

$$\Phi\left(\frac{\theta^* - \overline{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}\left(x^* - (\overline{\theta} + \bar{v}y)\right)}{\sqrt{\frac{(y^2\sigma_v^2 + \sigma_\varepsilon^2)\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}}}\right) = \frac{c}{\delta}. \tag{25}$$

This means that the citizen is indifferent between attacking and refraining from doing so if the observed signal x^* is such that the posterior probability of the attack being successful (i.e. of θ being below θ^*) is equal to $\frac{c}{\delta}$. With a few simple transformations, we obtain x^* explicitly:

$$x^* = \bar{\theta} + \bar{v}y + \left(\theta^* - \bar{\theta}\right)\frac{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}{\sigma_\theta^2} - \sqrt{\frac{\left(\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2\right)\left(y^2\sigma_v^2 + \sigma_\varepsilon^2\right)}{\sigma_\theta^2}}\Phi^{-1}\left(\frac{c}{\delta}\right). \tag{26}$$

Clearly, the critical private signal, x^*, is decreasing in $\frac{c}{\delta}$. The higher is the citizen's cost of attacking the regime, the higher must be the posterior probability that the attack will be successful for the citizen to be willing to attack, and so—as intuition would suggest—the lower must be the citizen's critical private signal, x^*, for her to choose to attack the regime. Furthermore, x^* is decreasing in $\bar{\theta}$: as the mean state of the world increases, the higher realisations of θ become more likely, and hence the highest value of the posterior which prompts the citizen to attack decreases. Finally, x^* is increasing in θ^*, which means that as the threshold for a successful attack increases—thus making it more likely that an attack will be successful—the citizen is willing to attack for higher and higher values of his posterior.

Given the variance of the private signal x, $Var(x) = \sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2$, we can now find the ex ante probability that there is no regime change:

$$\Phi\left(\frac{E(x) - x^*}{\sqrt{Var(x)}}\right) = \Phi\left(\left(\bar{\theta} - \theta^*\right)\frac{\sqrt{\sigma_\theta^2 + y^2\sigma_v^2 + \sigma_\varepsilon^2}}{\sigma_\theta^2} + \sqrt{\frac{y^2\sigma_v^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right). \tag{27}$$

Using (27) and subtracting the expected direct cost of employing online commentators, we can now state the government's ex ante utility as in Lemma 1. □

Lemma 3

Proof. If we partially differentiate the expression for an equilibrium value of the critical state of the world, θ^*, with respect to the mean state, $\bar{\theta}$, we obtain:

$$\frac{\partial\theta^*(y;\cdot)}{\partial\bar{\theta}} = \frac{-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}}{1-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}}. \tag{28}$$

The numerator is clearly negative. By the properties of the normal distribution, the maximum value which $\phi(\cdot)$ attains is $1/\sqrt{2\pi}$. Given the equilibrium uniqueness condition stated in Proposition 5, this ensures that the denominator is positive.

The government's benefit, $B_P(y;\cdot)$, is defined here as the probability that there is no regime change multiplied by the weighting parameter, μ. Using the expression for the government's benefit, $B_P(y;\cdot) := \mu\Phi\left(\left(\bar{\theta} - \theta^*\right)\left(\sigma_\theta^2\right)^{-\frac{1}{2}}\right)$, we obtain the following result:

$$\begin{aligned}\frac{\partial B_P(y;\cdot)}{\partial\bar{\theta}} &= -\frac{\mu}{\sqrt{\sigma_\theta^2}}\phi\left(\frac{\bar{\theta}-\theta^*}{\sqrt{\sigma_\theta^2}}\right)\frac{\partial\theta^*}{\partial\bar{\theta}} \\ &= \frac{\mu\phi\left(\frac{\bar{\theta}-\theta^*}{\sqrt{\sigma_\theta^2}}\right)\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}}{\sqrt{\sigma_\theta^2}\left(1-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\right)} > 0. \end{aligned} \tag{29}$$

□

Lemma 4

Proof. If we partially differentiate the expression for $\theta^*(y;\cdot)$ with respect to $\frac{c}{\delta}$, we obtain:

$$\frac{\partial\theta^*(y;\cdot)}{\partial\left(\frac{c}{\delta}\right)} = \frac{-\phi(\cdot)\sqrt{\frac{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\frac{\partial\Phi^{-1}\left(\frac{c}{\delta}\right)}{\partial\left(\frac{c}{\delta}\right)}}{1-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}}, \tag{30}$$

where $\phi(\cdot) = \phi\left(\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\left(\theta^*-\bar{\theta}\right)-\sqrt{\frac{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\Phi^{-1}\left(\frac{c}{\delta}\right)\right)$. 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 5. We can now use the expression for the government's benefit, $B_P(y;\cdot) := \mu\Phi\left(\left(\bar{\theta}-\theta^*\right)\left(\sigma_\theta^2\right)^{-\frac{1}{2}}\right)$, to obtain the following result:

$$\begin{aligned}\frac{\partial B_P(y;\cdot)}{\partial\left(\frac{c}{\delta}\right)} &= -\frac{\mu}{\sqrt{\sigma_\theta^2}}\phi\left(\frac{\bar{\theta}-\theta^*}{\sqrt{\sigma_\theta^2}}\right)\frac{\partial\theta^*}{\partial\left(\frac{c}{\delta}\right)} \\ &= \frac{\mu\phi\left(\frac{\bar{\theta}-\theta^*}{\sqrt{\sigma_\theta^2}}\right)\phi(\cdot)\sqrt{\frac{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}{\sigma_\theta^2}}\frac{\partial\Phi^{-1}\left(\frac{c}{\delta}\right)}{\partial\left(\frac{c}{\delta}\right)}}{\sqrt{\sigma_\theta^2}\left(1-\phi(\cdot)\frac{\sqrt{y^2\sigma_v^2+\sigma_\varepsilon^2}}{\sigma_\theta^2}\right)} > 0. \end{aligned} \tag{31}$$

□

Lemma 6

Case 1: Citizens' costs of attacking the regime are relatively high, $\frac{1}{2}\delta < c < \delta$.

We start the analysis of this case with an illustration of the threshold of marginal positive benefit:[31]

[31] Note here that the government's marginal benefit from increasing the online commentators' payment is positive (negative) whenever the critical state of the world, $\theta^*(y;\cdot)$ is decreasing (increasing) in the value of the payment, y.

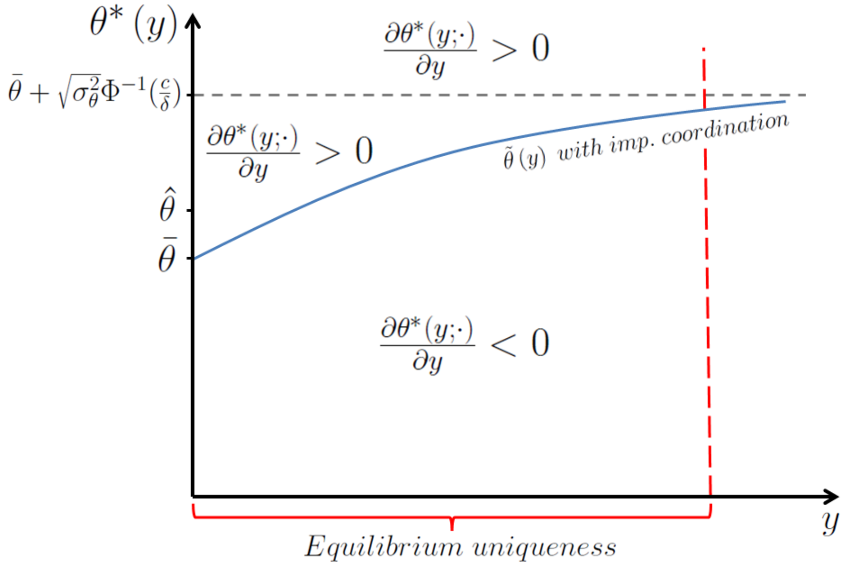


Figure 3. The impact of an increase in the online commentators payment, y, on the critical state of the world, $\theta^*(y;\cdot)$, when the citizens' costs of attacking the regime are relatively high, $\frac{1}{2}\delta < c < \delta$.

One interpretation of Figure 3 is that there is a tripartition of the space of citizens' sentiments. If the intrinsic fragility of the regime, $\theta^*(0;\cdot)$, is high relative to the mean state of the world, then the probability of regime change is increasing in the online commentators' payment, y, for all values of y.[32] Equivalently, this means that the government's benefit is unequivocally decreasing in y. On the other hand, if the intrinsic fragility of the regime is low relative to the mean state of the world, the probability of regime change is unequivocally decreasing in y. Finally, if the intrinsic fragility is intermediate, then the probability of regime change is increasing in y for small values of the payment, but then it is decreasing in y once the online commentators' payment reaches a certain level. We make the following claim:

Claim 1. *Suppose that the citizens' costs of attacking are relatively high, $\frac{1}{2}\delta < c < \delta$, and that the equilibrium uniqueness condition in Proposition 5 is satisfied. Then there exists a certain level of intrinsic fragility of the regime, $\tau \in \left[\bar{\theta}, \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})\right]$, such that:*

(i) If the intrinsic fragility is high enough, $\theta^*(0;\cdot) \geq \tau$, the lowest possible payment to online commentators, $y = 0$, maximises the government's benefit;

(ii) If the intrinsic fragility is low enough, $\theta^*(0;\cdot) \leq \tau$, the highest possible payment to online commentators, $y = y_H$, maximises the government's benefit.

[32] As already stated earlier, we are restricting the analysis to the values of the payment that are permitted by the equilibrium uniqueness condition from Proposition 5.

We start the proof of this simple cut-off result with a lemma:

Lemma A1. *Suppose that the cost of participating in an attack is relatively high,* $\frac{1}{2}\delta < c < \delta$, *and that the equilibrium uniqueness condition is satisfied. Then there exists a certain level of intrinsic fragility of the regime,* $\hat{\theta} \in \left[\bar{\theta}, \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})\right]$, *such that, if the intrinsic fragility is greater than* $\hat{\theta}$, *the probability of regime change is increasing in the online commentators' payment,* y, *for all values of* y.

Proof. Let us denote $\tilde{\theta}\,(y;\cdot) = \bar{\theta} + \sqrt{\frac{\sigma_\theta^2(y^2\sigma_v^2+\sigma_\varepsilon^2)}{\sigma_\theta^2+y^2\sigma_v^2+\sigma_\varepsilon^2}}\Phi^{-1}(\frac{c}{\delta})$, where $\tilde{\theta}\,(y;\cdot)$ is the threshold below which $\theta^*\,(y;\cdot)$ is decreasing in y, and $y_H = \left[0, \sqrt{\left(2\pi\left(\sigma_\theta^2\right)^2 - \sigma_\varepsilon^2\right)/\sigma_v^2}\right]$. Note that $\tilde{\theta}\,(y;\cdot)$ is strictly increasing in y, with an upper bound of $\bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})$. Thus, the condition that $\theta^*(0) \geq \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})$ ensures that $\frac{\partial\theta^*}{\partial y} > 0$ for all $y \in [0, y_H]$. This means that the government's benefit is certainly decreasing in y for all $y \in [0, y_H]$ when $\theta^*(0) \geq \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})$.

Now, analyse the case where $\bar{\theta} < \theta^*\,(0) < \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})$. Note that the derivative $\frac{\partial\theta^*}{\partial y}$ is positive for small y. If $\theta^*\,(0)$ is high enough, then as we increase y, $\theta^*\,(y;\cdot)$ will reach $\bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})$ and so there will be no $\hat{y}$ such that $\theta^*\,(\hat{y};\cdot) = \tilde{\theta}\,(\hat{y};\cdot)$. As a result, $\theta^*\,(y;\cdot)$ will be increasing in y for all $y \in [0, y_H]$. On the other hand, if $\theta^*(0)$ is low enough, then there will be $\hat{y} < y_H$ such that $\theta^*\,(\hat{y};\cdot) = \tilde{\theta}\,(\hat{y};\cdot)$, and $\theta^*\,(\hat{y};\cdot)$ will be increasing in y for $y < \hat{y}$ and decreasing in y for $y > \hat{y}$. □

It follows from Lemma A1 that the government optimally chooses the lowest possible payment whenever the intrinsic fragility of the regime, $\theta^*\,(0;\cdot)$, is greater than $\hat{\theta}$. Furthermore, note that when the intrinsic fragility is less than $\bar{\theta}$, an increase in the payment, y, leads to a decrease in the critical state of the world, $\theta^*\,(y;\cdot)$, for all values of y. Thus, whenever the intrinsic fragility is so low, setting the highest possible payment minimises the probability of a regime change and, thereby, it also maximises the government's benefit.

To complete the proof of Claim 1, we still have to consider the interval where the intrinsic fragility is intermediate: $\bar{\theta} \leq \theta^*\,(0;\cdot) \leq \hat{\theta}$. Within this interval, the critical state of the world, $\theta^*\,(y;\cdot)$, is strictly increasing in y for low values of the online commentators' payment, and is decreasing in y once the payment becomes large enough. Clearly, the probability of regime change is then minimised for an extreme value of the payment, i.e. the government will choose either the lowest possible payment, $y = 0$, or the highest possible, $y = y_H$. If the intrinsic fragility is in the lower part of this interval, then $\theta^*\,(y;\cdot)$ touches the $\tilde{\theta}\,(y;\cdot)$ threshold (below which $\theta^*\,(y;\cdot)$ is decreasing in y) for a relatively low value of y, and setting the payment as high as possible allows the government to minimise the probability of regime change. On the other hand, if the intrinsic fragility is in the upper part of the interval, then a reverse argument holds, and the lowest possible payment minimises the probability of regime change.

It follows that there is a threshold, $\tau \in \left[\bar{\theta}, \bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})\right]$, such that $y = y_H$ minimises the probability of regime change whenever $\theta^*(0; \cdot) \leq \tau$, whereas $y = 0$ minimises the probability of regime change if the inequality is reversed.

Case 2: Citizens' costs of attacking the regime are intermediate, $c = \frac{1}{2}\delta$.

Again, we first illustrate the threshold of positive marginal benefit:

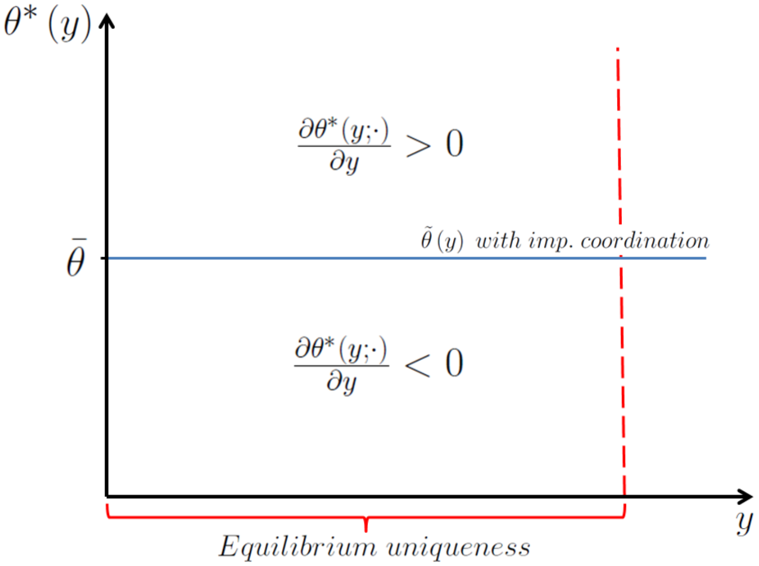

$\theta^*(y)$
$\frac{\partial\theta^*(y;\cdot)}{\partial y} > 0$
$\bar{\theta}$
$\tilde{\theta}(y)$ with imp. coordination
$\frac{\partial\theta^*(y;\cdot)}{\partial y} < 0$
y
Equilibrium uniqueness

Figure 4. The impact of an increase in the online commentators payment, y, on the critical state of the world, $\theta^*(y; \cdot)$, when the citizens' costs of attacking the regime are intermediate, $c = \frac{1}{2}\delta$.

In this knife-edge intermediate case, we observe a simple cut-off result:

Claim 2. *Suppose that the citizens' costs of attacking are intermediate, $c = \frac{1}{2}\delta$, and that the equilibrium uniqueness condition in Proposition 5 is satisfied. Then:*

(i) If the intrinsic fragility is higher than the mean state, $\theta^*(0; \cdot) > \bar{\theta}$, the lowest possible payment to online commentators, $y = 0$, maximises the government's benefit;

(ii) If the intrinsic fragility is equal to the mean state, $\theta^*(0; \cdot) = \bar{\theta}$, the government's benefit is the same for all values of online commentators' payment;

(iii) If the intrinsic fragility is lower than the mean state, $\theta^*(0; \cdot) < \bar{\theta}$, the highest possible payment to online commentators, $y = y_H$, maximises the government's benefit.

Lemma 5 implies here that the threshold of positive marginal benefit, $\tilde{\theta}(y)$, is independent of the value of the payment, y. Hence, whenever the intrinsic fragility of the regime, $\theta^*(0;\cdot)$, is greater (smaller) than the mean state of the world, $\bar{\theta}$, the probability of regime change is increasing (decreasing) in the online commentators' payment, y, for all values of y. Then it is straightforward to observe the statements (i)-(iii) of Proposition 2.

Case 3: Citizens' costs of attacking the regime are relatively low, $c < \frac{1}{2}\delta$.

Again, we start the analysis by illustrating the threshold of marginal positive benefit:

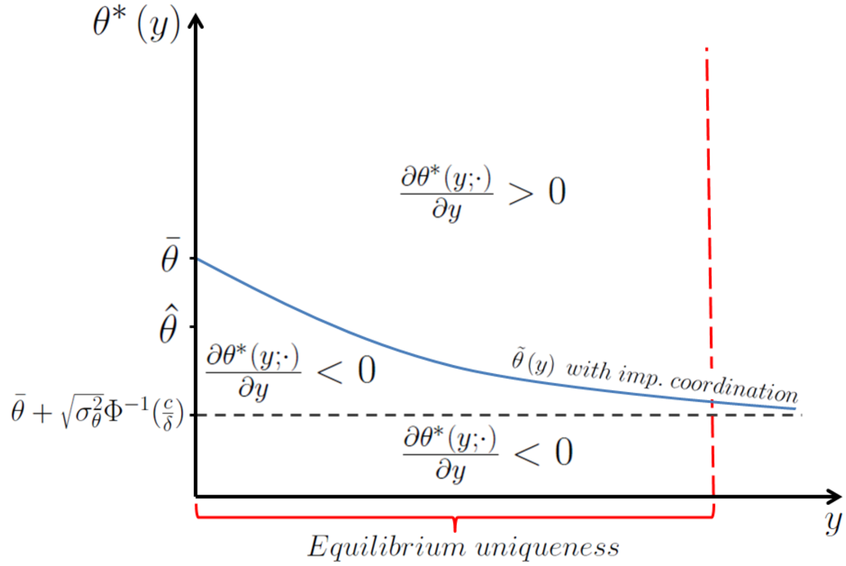

$\theta^*(y)$
$\frac{\partial\theta^*(y;\cdot)}{\partial y} > 0$
$\bar{\theta}$
$\hat{\theta}$
$\frac{\partial\theta^*(y;\cdot)}{\partial y} < 0$
$\tilde{\theta}(y)$ with imp. coordination
$\bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta})$
$\frac{\partial\theta^*(y;\cdot)}{\partial y} < 0$
y
Equilibrium uniqueness

Figure 5. The impact of an increase in the online commentators payment, y, on the critical state of the world, $\theta^*(y;\cdot)$, when the citizens' costs of attacking the regime are relatively low, $c < \frac{1}{2}\delta$.

Similarly to the previous two cases, the government's benefit is maximised by setting the online commentators' payment as high as possible if the intrinsic fragility of the regime, $\theta^*(0;\cdot)$, is low enough. Furthermore, when the intrinsic fragility is high enough, the government maximises its benefit by setting the payment at zero. However, there is also an important difference relative to the previous two cases: the value of the online commentators' payment that maximises the government's benefit may now be an interior solution. The following proposition states the results formally:

Claim 3. *Suppose that the citizens' costs of attacking are relatively low, $c < \frac{1}{2}\delta$, and that the equilibrium uniqueness condition in Proposition 5 is satisfied. Then there exists a certain level of intrinsic fragility of the regime, $\tau' \in \left[\bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta}), \bar{\theta}\right]$, such that:*

(i) If the intrinsic fragility is higher than the mean state, $\theta^*(0;\cdot) > \bar{\theta}$, the lowest possible payment to online commentators, $y = 0$, maximises the government's benefit;

(ii) If the intrinsic fragility is lower than the mean state but high enough, $\tau' \leq \theta^*(0;\cdot) \leq \bar{\theta}$, the government's benefit is maximised for a value of the payment which is within the interval $(0, y_H)$ and is decreasing in $\theta^*(0;\cdot)$;

(iii) If the intrinsic fragility is low enough, $\theta^*(0;\cdot) < \tau'$, the highest possible payment to online commentators, $y = y_H$, maximises the government's benefit.

We start the proof of this proposition with the following lemma:

Lemma A2. *Suppose that the cost of participating in an attack is relatively low, $c < \frac{1}{2}\delta$, and that the equilibrium uniqueness condition is satisfied. Then there exists a certain level of intrinsic fragility of the regime, $\tau' \in \left[\bar{\theta} + \sqrt{\sigma_\theta^2}\Phi^{-1}(\frac{c}{\delta}), \bar{\theta}\right]$, such that, if $\theta^*(0;\cdot) < \tau'$, the government's benefit is increasing in the online commentators' payment, y, for all values of y.*

The proof of Lemma A2 follows steps that are analogous to the proof of Lemma A1. Lemma A2 implies here that the government optimally sets the highest possible payment, $y = y_H$, whenever $\theta^*(0;\cdot) \leq \tau'$. Moreover, note that when the intrinsic fragility is higher than the mean state, $\theta^*(0;\cdot) > \bar{\theta}$, an increase in the payment leads to a decrease in the critical state, $\theta^*(y;\cdot)$, for all values of y. In that case, choosing the lowest possible payment minimises the probability of regime change and, thus, it also maximises the government's benefit.

However, to complete the proof of Claim 3, we still have to consider the intermediate interval, $\tau' \leq \theta^*(0;\cdot) \leq \bar{\theta}$. Within this interval, the critical state, $\theta^*(y;\cdot)$, is strictly decreasing in y for low values of the online commentators' payment, and is strictly increasing in y once the payment becomes large enough. Thus, the probability of regime change will be minimised for the value of the payment which solves $\theta^*(y^*;\cdot) = \tilde{\theta}(y^*;\cdot)$, where $\tilde{\theta}(y;\cdot)$ is the threshold of positive marginal benefit. This yields part (ii) of Claim 3.

Lemma 8

Proof. The first order condition implies that $\mu\bar{v} = \left(\bar{\theta} + 2\bar{v}y\right)(1 + y^2\eta + \xi)$. Thus, we have the following polynomial:

$$f(y^*;\cdot) = 2\,(y^*)\,^3\bar{v}\eta + (y^*)\,^2\bar{\theta}\eta + 2y^*\bar{v}(1+\xi) + \bar{\theta}(1+\xi) - \mu\bar{v} = 0. \tag{32}$$

Implicit differentiation of y^* with respect to $\bar{v}$ yields:

$$\frac{\partial y^*}{\partial \bar{v}} = \frac{-\frac{\partial f}{\partial \bar{v}}}{\frac{\partial f}{\partial y^*}} = -\frac{2\,(y^*)\,^3\eta + 2y^*(1+\xi) - \mu}{6\,(y^*)\,^2\bar{v}\eta + 2y^*\bar{\theta}\eta + 2\bar{v}(1+\xi)}. \tag{33}$$

Since the denominator is positive for any value of $y^* \geq 0$, the following property holds: $\frac{\partial y^*}{\partial v} > 0 \iff g(y^*) = (y^*)^3 + \frac{1+\xi}{\eta} y^* - \frac{\mu}{2\eta} < 0$. Clearly, the function $g(y^*)$ has only one real root (the function $g(y^*)$ is strictly increasing), and this root is always positive (since $g(0) = -\frac{\mu}{\eta} < 0$). □

7.2 Relationship of Section 5 with Models of Career Concerns

The linear model with the payment not observed by the citizens is closely related to models of managerial career concerns, in particular Dewatripont, Jewitt, and Tirole (1999a, 1999b). Applying their notation to our model, let $f(\theta, x \mid y)$ denote the joint density of the state of the world ("talent" in the language of economics of career concerns) and observables given the payment y ("effort"). The marginal density of the observables is:

$$\hat{f}(x \mid y) = \int f(\theta, x \mid y)\, d\theta. \tag{34}$$

The government's reward for equilibrium "effort", y^*, and the (commentator's) "performance", x, is therefore:

$$t = E(\theta \mid x, y^*) = \int \theta \frac{f(\theta, x \mid y^*)}{\hat{f}(x \mid y^*)} d\theta. \tag{35}$$

Also, like in the companion papers by Dewatripont et al., let c_y and $\hat{f}_y$ denote the gradients with respect to effort of the cost function and the marginal distribution. Suppose that the citizens anticipate equilibrium payment y^*. The government then chooses payment y^* to maximise its expected utility (since only the general distributions of θ, v, and ε are common knowledge):

$$\max \mathbb{E}\left[\mathbb{E}\left[\theta \mid x, y^*\right]\right] - c(y) \tag{36}$$

Dewatripont et al. (see Proposition 2.1 in both companion papers) show that this yields the equivalent of the following result (with different labelling of variables):

Proposition 2.1. *In an equilibrium, the gradient of the cost function is equal to the covariance of the state of the world and the likelihood ratio:*

$$Cov\left(\theta, \frac{\hat{f}_y}{\hat{f}}\right) = c_y(y^*). \tag{37}$$

Dewatripont et al. (1999b) also analyse a single-task normal example as an illustration of their Proposition 2.1. However, unlike in their paper, in our model there is also a random variable in front of the "effort" variable. This means we cannot transform their normal example into our model via relabelling. Nevertheless, we can use their Proposition 2.1 to confirm our

Proposition 8.

First, note that $\hat{f}\left(x \mid y\right)$ is proportional to $exp\left[-\frac{\left(x-\bar{\theta}-\bar{v}y\right)^{2}}{2\left(\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)}\right]$. The gradient of the marginal distribution with respect to effort, $\hat{f}_{y}\left(x \mid y\right)$, is proportional to:

$$\begin{aligned}\hat{f}_{y}\left(x \mid y\right) &\propto exp\left[-\frac{\left(x-\bar{\theta}-\bar{v}y\right)^{2}}{2\left(\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right)}\right]\left[\frac{\bar{v}\left(x-\bar{\theta}-\bar{v}y\right)Var\left(x\right)+y\sigma_{v}^{2}\left(x-\bar{\theta}-\bar{v}y\right)^{2}}{\left[Var\left(x\right)\right]^{2}}\right] \\ &= exp\left[\cdot\right]\left[\frac{\bar{v}\left(x-E\left(x\right)\right)Var\left(x\right)+y\sigma_{v}^{2}\left(x-E\left(x\right)\right)^{2}}{\left[Var\left(x\right)\right]^{2}}\right], \qquad (38)\end{aligned}$$

where $Var\left(x\right)=\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}$, and $x-E\left(x\right)=\left(\theta-\bar{\theta}\right)+\left(v-\bar{v}\right)y+\varepsilon$. As an intermediate step in calculating $Cov\left(\theta,\frac{\hat{f}_{a}}{\hat{f}}\right)$, note that:

$$\begin{aligned}Cov\left(\theta,\left(\theta-\bar{\theta}\right)^{2}\right) &= E\left[\theta\left(\theta-\bar{\theta}\right)^{2}\right]-E\left[\theta\right]E\left[\left(\theta-\bar{\theta}\right)^{2}\right] \\ &= \underbrace{E\left[\theta\left(\theta-\bar{\theta}\right)^{2}\right]-E\left[\bar{\theta}\left(\theta-\bar{\theta}\right)^{2}\right]}_{skewness=0} \\ &\quad \underbrace{+E\left[\bar{\theta}\left(\theta-\bar{\theta}\right)^{2}\right]-E\left[\theta\right]E\left[\left(\theta-\bar{\theta}\right)^{2}\right]}_{\equiv 0} \qquad (39) \\ &= 0. \qquad (40)\end{aligned}$$

Thus:

$$\begin{aligned}Cov\left(\theta,\frac{\hat{f}_{a}}{\hat{f}}\right) &= Cov\left(\theta,\frac{\bar{v}\left(x-E\left(x\right)\right)Var\left(x\right)+y\sigma_{v}^{2}\left(x-E\left(x\right)\right)^{2}}{\left[Var\left(x\right)\right]^{2}}\right) \\ &= Cov\left(\theta,\frac{\bar{v}\left(x-E\left(x\right)\right)}{Var\left(x\right)}\right) \\ &= \frac{\sigma_{\theta}^{2}\bar{v}}{\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}}, \qquad (41)\end{aligned}$$

where, as before, $Var\left(x\right)=\sigma_{\theta}^{2}+y^{2}\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}$, and $x-E\left(x\right)=\left(\theta-\bar{\theta}\right)+\left(v-\bar{v}\right)y+\varepsilon$. We now need to multiply this expression for covariance by μ to obtain the marginal benefit of the government. Equating this with the marginal cost of increasing y, which is equal to $\bar{\theta}+2\bar{v}y$, yields equation (**??**).

References

[1] **Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan**, "Dynamic Global Games of Regime Change: Learning, Multiplicity, and the Timing of Attacks," *Econometrica*, May 2007, 75(3): 711-756.

[2] **Bannier, Christina E., and Frank Heinemann**, "Optimal Transparency and Risk-Taking to Avoid Currency Crises," *Journal of Institutional and Theoretical Economics*, 2005, 161(3): 374-391.

[3] **Bénabou, Roland, and Jean Tirole**, "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, Jul. 2003, 70(3): 489-520.

[4] **Bénabou, Roland, and Jean Tirole**, "Incentives and Prosocial Behavior," *American Economic Review*, Dec. 2006, 96(5): 1652-1678.

[5] **Besley, Timothy, and Andrea Prat**, "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability," *American Economic Review*, Jun. 2006, 96(3): 720-736.

[6] **Boix, Carles, and Milan W. Svolik**, "The Foundations of Limited Authoritarian Government: Institutions and Power Sharing in Dictatorships," *Journal of Politics*, Apr. 2013, 75(2): 300–316.

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

[8] **Carlsson, Hans, and Eric van Damme**, "Global Games and Equilibrium Selection," *Econometrica*, Sep. 1993, 61(5): 989-1018.

[9] **Chassang, Sylvain, and Gerard Padró-i-Miquel**, "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics*, Nov. 2010, 125(4): 1821-1858.

[10] **Dellarocas, Chrysanthos**, "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science*, Oct. 2006, 52(10): 1577-1593.

[11] **Dewatripont, Mathias, Ian Jewitt, and Jean Tirole**, "The Economics of Career Concerns, Part I: Comparing Information Structures," *Review of Economic Studies*, Jan. 1999, 66(1): 183-198.

[12] **Dewatripont, Mathias, Ian Jewitt, and Jean Tirole**, "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies," *Review of Economic Studies*, Jan. 1999, 66(1): 199-217.

[13] **Edmond, Chris**, "Information Manipulation, Coordination, and Regime Change," *Review of Economic Studies*, Oct. 2013, 80(4): 1422-1458.

[14] **Egorov, Georgy, Sergei Guriev, and Konstantin Sonin**, "Why Resource-poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data," *American Political Science Review*, Nov. 2009, 103(4): 645-668.

[15] **Gehlbach, Scott, and Konstanin Sonin**, "Government Control of the Media," *Journal of Public Economics*, Oct. 2014, 118: 163-171.

[16] **Gneezy, Uri, and Aldo Rustichini**, "A Fine Is a Price," *Journal of Legal Studies*, Jan. 2000, 29(1): 1–17.

[17] **Guriev, Sergei, and Daniel Treisman**, "How Modern Dictators Survive: Cooptation, Censorship, Propaganda, and Repression," CEPR Discussion Paper No. 10454, March 2015.

[18] **Han, Rongbin**, "Manufacturing Consent in Censored Cyberspace: State-Sponsored Online Commentators on Chinese Internet Forums," APSA 2012 Annual Meeting Paper, 2012.

[19] **Hellwig, Christian**, "Public Information, Private Information, and the Multiplicity of Equilibria in Coordination Games," *Journal of Economic Theory*, Dec. 2002, 107(2): 191-222.

[20] **Holmström, Bengt**, "Managerial Incentive Problems - A Dynamic Perspective," *Review of Economic Studies*, Jan. 1999, 66(1): 169-182.

[21] **Holmström, Bengt, and Paul Milgrom**, "The Firm as an Incentive System," *American Economic Review*, Sep. 1994, 84(4): 972-991.

[22] **King, Gary, Jennifer Pan, and Margaret E. Roberts**, "How Censorship in China Allows Government Criticism but Silences Collective Expression," *American Political Science Review*, May 2013, 107(2): 1-18.

[23] **Loeper, Antoine, Jakub Steiner, and Colin Stewart**, "Influential Opinion Leaders," *Economic Journal*, Dec. 2014, 124(581): 1147-1167.

[24] **Lorentzen, Peter**, "China's Strategic Censorship," *American Journal of Political Science*, Apr. 2014, 58(2): 402-414.

[25] **Mayzlin, Dina, Yaniv Dover, and Judith Chevalier**, "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, Aug. 2014, 104(8): 2421-55.

[26] **Metz, Christina E.**, “Private and Public Information in Self-Fulfilling Currency Crises,” *Journal of Economics*, 2002, 76(1): 65-85.

[27] **Morris, Stephen, and Hyun S. Shin**, “Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks,” *American Economic Review*, Jun. 1998, 88(3): 587–597.

[28] **Morris, Stephen, and Hyun S. Shin**, “Rethinking Multiple Equilibria in Macroeconomics,” in NBER Macroeconomics Annual 2000, ed. by B. S. Bernanke and K. S. Rogoff, Cambridge, MA: MIT Press, 2001, pp. 139-161.

[29] **Morris, Stephen, and Hyun S. Shin**, “Global Games—Theory and Applications,” in “Advances in Economics and Econometrics,” 8th World Congress of the Econometric Society, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky, Cambridge, U.K.: Cambridge University Press, 2003, pp. 56–114.

[30] **Shadmehr, Mehdi, and Dan Bernhardt**, “Collective Action with Uncertain Payoffs: Coordination, Public Signals and Punishment Dilemmas,” *American Political Science Review*, Nov. 2011, 105(4): 829–851.

[31] **Shadmehr, Mehdi, and Dan Bernhardt**, “State Censorship,” *American Economic Journal: Microeconomics*, 2015, forthcoming.

[32] **Titmuss, Richard**, “The Gift Relationship: From Human Blood to Social Policy,” London: Allen and Unwin, 1970.