Global Manipulation by Local Obfuscation [*]

Fei Li[†] Yangbo Song[‡] Mofei Zhao[§]

December 5, 2019

Abstract

We study information design in a regime change context. A continuum of agents simultaneously choose whether to attack the current regime and will succeed if and only if the mass of attackers outweighs the regime's strength. A designer manipulates information about the regime's strength to maintain the status quo. The optimal information structure exhibits local obfuscation, some agents receive a signal matching the true strength of the status quo, and others receive an elevated signal professing slightly higher strength. Public signals are strictly suboptimal, and in some cases where public signals become futile, local obfuscation guarantees the status quo's survival.

Keywords: Bayesian persuasion, coordination, information design, obfuscation, regime change

JEL Classification: C7, D7, D8.

[*] We thank Yu Awaya, Arjada Bardhi, Gary Biglaiser, Daniel Bernhardt, James Best, Jimmy Chan, Yi-Chun Chen, Toomas Hinnosaar, Ju Hu, Tetsuya Hoshino, Chong Huang, Nicolas Inostroza, Qingmin Liu, George Mailath, Laurent Mathevet, Stephen Morris, Peter Norman, Mallesh Pai, Alessandro Pavan, Jacopo Perego, Xianwen Shi, Joel Sobel, Satoru Takahashi, Can Tian, Kyle Woodward, Xi Weng, and Jidong Zhou for comments.

[†] Department of Economics, University of North Carolina, Chapel Hill. Email: lifei@email.unc.edu.

[‡] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen. Email: yangbosong@cuhk.edu.cn.

[§] School of Economics and Management, Beihang University, Beijing. Email: zhaomf.06@gmail.com.

> *"The conscious and intelligent manipulation of the organized habits and opinions of the masses is an important element in democratic society. Those who manipulate this unseen mechanism of society constitute an invisible government which is the true ruling power of our country."*
>
> — Edward Bernays

1 Introduction

The revolution of information and communication technology raises growing concerns about digital authoritarianism.[1] Governments worldwide are increasingly using advanced technology to consolidate their control over information and suppress dissent.[2] Despite their tremendous effort to uphold the notion of "cyber sovereignty," to establish information censorship, and to spread disinformation, full manipulation of information in the modern age remains outside the grasp of autocrats. To optimistic liberals, it keeps the hope alive. This paper begins the formal investigation of whether this hope is justified. We study information design in a canonical regime-change game *à la* Morris and Shin (2003) and derive the optimal information structure to stabilize the regime. The result paints a bleak picture: the most effective policy to maintain the status quo requires only creating a minor obfuscation for a fraction of agents.

In our model, an information designer faces a unit mass of agents who simultaneously decide whether to coordinate on an attack. Attacking is costly, and each attacker will be rewarded if and only if the status quo is overthrown. The strength of the status quo, namely the state, is randomly selected from an interval by nature and unknown to the agents. The status quo survives if and only if the total measure of attackers does not exceed its state. When the state is above one, it is *invincible* because the status quo survives under the attack of all agents. The information designer commits to a state-dependent information policy that sends a signal, which can be public or private, to each agent, and his objective is to maximize the regime's probability of survival in its least-preferred equilibrium among agents.

It is helpful to first examine a few benchmark cases. First, if the information designer always reveals the state publicly, the agents will attack, and the regime, in turn, survives, if and only if the state is invincible. Contrarily, if the information designer reveals no information, the agents either always or never coordinate on attacking, depending on their prior, and the regime survives under attacks if and only if its state is invincible. Finally, if the information designer is constrained to using public information policies, an explicitly solvable state cutoff exists where at optimum the infor-

[1] See Yuval Noah Harari, "Why Technology Favors Tyranny," *The Atlantic*, October 322(3), 2018.

[2] See Frenkel, Sheera, Kate Conger, and Kevin Roose. "Russia's playbook for social media disinformation has gone global." *The New York Times* 31 (2019).

mation designer sends one distinct signal for states above this cutoff and another for those below. The agents attack upon seeing the second signal.

The main contribution of this paper is an explicit and straightforward characterization of the optimal information structure. Its simplest implementation features a countable set of messages — regardless of the state set's countability — and private i.i.d. signal. At optimum, the information designer partitions the state set by a strength threshold. For weak states below the threshold, a determinate self-identifying signal is sent, which will coordinate all agents on attacking. Among the strong states above the threshold, the information designer further classifies them into tiers by strength. The first tier, i.e. the set of the strongest ones, sends signal s_1 to all agents; the second tier sends s_1 to a proportion of agents, and sends another signal s_2 to others; the third tier sends s_2 to a proportion of agents, and sends another signal s_3 to others; and so on. In other words, except the first tier and the weak states which will always be overthrown, the information designer under each state executes a "truth-lie" policy, essentially revealing its tier to some agents but deceiving other agents by a slightly stronger tier. In this way, the information structure generates *local obfuscation* among agents.

The optimal information structure collapses *global coordination* among agents by creating a hierarchy of private beliefs. Under the optimal information policy, an agent seeing a tier-1 signal will refrain from attacking since the probability of facing an invincible status quo is sufficiently high; realizing this, she will not attack upon a tier-2 signal either, because a fraction of her fellow agents are likely to receive a tier-1 signal and be deterred, resulting in an insufficient mass of attackers in expectation. The use of a tier-1 signal, therefore, generates a ripple effect to squelch agents' attack on lower-tier signals. The iteration then unravels to the limit case that no single agent seeing a signal sent by any tier will choose to attack, and thus all states beyond the above threshold survive. It is worth noting that, although each state in the optimal design needs only two signals, the maximized survival probability is not achievable with a (public or private) disclosure policy of binary recommendations, because the iteration process among agents can only result from endogenously heterogeneous posterior beliefs and higher-order uncertainty.

Local obfuscation has a unique advantage over public information structures that send each agent the same signal conditional on the state. We demonstrate this advantage in two distinct ways. On the one hand, when the measure of invincible states converges to zero, the optimal public information structure becomes futile, while optimal local obfuscation still manages to save a significant measure of states below one. A sharp implication of this result is that when the attacking cost is sufficiently high but the measure of invincible states becomes almost negligible, virtually no state survives under public information disclosure, but all states survive under optimal local obfuscation. In the latter case, the information designer is even relieved of the usual *commitment* concern because the survival of the status quo is guaranteed after all.[3] On the other hand, given a target set of surviving states, optimal local obfuscation allows for a lower threshold of attacking cost to achieve the target than optimal public disclosure, and the difference between the cost thresholds coincides with the conditionally expected strength of the surviving states below one.

The optimal design with local obfuscation can be easily implemented by a private i.i.d. signal. A natural and practical interpretation of this is that the information designer, being, for instance, the propaganda department of an autocratic regime, spreads messages with close but distinct implications across various social media. The designer does not need to control over the identity of receivers of each message, but it manages to *randomly* sow confusion and doubt, resulting in local obfuscation by dividing message recipients by media coverage.[4]

Besides straightforward implementation, optimal local obfuscation is robust in several ways. First, the identification of the threshold of surviving states, and thus the maximum survival probability of the status quo, results from the converging iterated process and does not require any prior knowledge of the optimum. Besides, when the information designer has only limited available signal values at her disposal, a natural truncation of our information structure that encompasses the highest tiers proves to be the unique optimum. Finally, this information structure remains optimal when arbitrary correlation across signals is allowed.

Related literature. Our paper belongs to the growing literature on Bayesian persuasion and information design initiated by Kamenica and Gentzkow (2011), and Bergemann and Morris (2016). The backbone of our framework is a game of regime change *à la* Morris and Shin (2003),[5] and the key novelty of our modeling approach is to fully incorporate agents' signals in the information design. On the front of information design in a global game context, the models closest to ours are Goldstein and Huang (2016, 2018), and Inostroza and Pavan (2018). They show if the design were to restrict attention to public information policies, the optimal one would be a simple monotone pass/fail test. Basak and Zhou (2018, 2019) consider such persuasion under dynamic settings. In most of these models, the agents' private signal assumes an exogenous noise, and another simple public signal is added in as the medium of persuasion; while in our model we directly look for the optimal design of the distribution of the private

[3] In general, the commitment assumption is justified by long-term interaction (Best and Quigley 2017; Mathevet et al. 2018) and restrictions for designers to fabricate the outcome of the information structure. (Guo and Shmaya 2019; Lipnowski et al. 2018).

[4] It is an open secret that these strategies are widely employed in many authoritarian regimes. For example, King et al. (2017) empirically study online comments by the notorious "50-cent gang" who post fabricated social media comments as if they were the genuine opinions of ordinary citizens. Also, Ong and Cabanes (2018) report interviews of architects of disinformation, the high-level strategists and digital support workers, behind fake news and "digital black operation" in Philippines.

[5] Also see Carlsson and van Damme (1993), Morris and Shin (1998), Jeitschko and Taylor (2001), Chassang and Miquel (2010), Shadmehr and Bernhardt (2011), Huang (2017), Cong et al. (2019), and Dai and Yang (2018) for other applications of coordination games.

signal. One exception is Inostroza and Pavan (2018), who allow the designer to communicate with agents privately, but private signals are constrained to be Gaussian in an extension.

Our paper assumes that the designer anticipates the adversarial (or worst) equilibrium to him being played for each information structure. This adversarial equilibrium selection is typical to study information design in coordination games. However, because the standard revelation principle argument does not apply, solving the optimal information structure with adversarial equilibrium selection is difficult and often relies on the structure of the base game. Recently, two methodological breakthroughs have been made. Mathevet et al. (2019) propose a belief-based method to analyze information design in finite games, which allows for flexible equilibrium selection. Morris et al. (2019) use the Bayes-correlated equilibrium approach to study adversarial information design in finite super-modular games where players choose between two actions. They provide a general and complete characterization of adversarial equilibrium implementable outcomes and sufficient conditions for the "perfect coordination property"– all players all choose the same action at the optimal information structure. We differ from these two papers by studying a base game with continuum states and agents, and our focus is to characterize an intuitive optimal information structure and highlight local obfuscation. It is worth mentioning that our optimal information design leads to strategic uncertainty and overlapping information sets among agents. This is similar to the exogenous information structure in the email game of Rubinstein (1989) and the follow-up literature on higher-order beliefs such as Carlsson and van Damme (1993), Morris et al. (1995) and Kajii and Morris (1997). Our paper complements this literature by endogenizing such information structures. See Bergemann and Morris (2019) and Morris et al. (2019) for more discussion on the connection between adversarial information design and the literature on higher-order beliefs.

Information design with multiple audiences has been studied in other settings, and the advantage of using private signals has been pointed out. Alonso and Câmara (2016), Bardhi and Guo (2018), Chan et al. (2019) analyze information design in voting models. Hoshino (2019) shows that, for any non-degenerate prior, agents can be persuaded to take an action profile which satisfies a generalization of risk dominance. Galperti and Perego (2018) formulate the multi-agent information design problem as a linear program, examine its dual representation, and provide a general characterization. Galperti and Perego (2019) investigate an information design problem where receivers share information with each other on a network of social links. Ely (2017) studies information design in a dynamic two-agent coordination game. In this literature, the outstanding performance of discriminatory information structure typically requires the designer to manage the statistical correlation between target signals of agents. On the contrary, the optimal information structure in the current paper is completely anonymous.

This paper also contributes to the literature on information manipulation in revolutions and political regime shifts. Our framework is closely related to the coordination game of regime change by Angeletos et al. (2006) and Edmond (2013).[6] These two papers also consider an informed government that endogenizes information towards a large number of imperfectly informed agents, but specifies different strategies for the government's information manipulation and provide distinct insights on the multiplicity of equilibria. Our main contrast to them is that while their models feature costly signaling without commitment, ours focuses on persuasion as the information designer can commit to the information release policy.

Also related to this paper is the literature on rumors in economics and other fields of social sciences. When an information designer sends mixed signals, the informational effect is comparable to spreading rumors, for example in Chen et al. (2016). Unlike their model, which emphasizes how agents communicate about rumors, and other approaches such as Banerjee (1993), which models rumors as a dynamic transmission process, we focus on the information designer's choice of whether to keep information obfuscated, and show that the information designer sometimes benefits from deliberate truthfulness.

More broadly, our paper is related to the literature about endogenous polarization. Glaeser (2005) provides a theory where politicians create hatred by spreading false stories about other groups' crimes to gain political support. Alder and Wang (2019) argue that political elites have the incentive to create mistrust among different social groups to "divide and rule." Jeong (2019) points out that a group of agents with private information about multi-dimensional states can be polarized by a public message.

Organization. The rest of the paper is organized as follows. Section 2 lays out the model. Section 3 presents the main result, which explicitly characterizes the optimal information structure, and compares optimal local obfuscation with optimal public propaganda. Section 4 examines the robustness of our proposed information structure. Section 5 concludes. Proofs, unless otherwise specified, are in the Appendices.

2 Model

Base game. The society is populated by a unit mass of agents, indexed by $i \in [0,1]$. There are two possible regimes, the status quo, and an alternative. Agent i decides to attack the current regime $(a_i = 1)$ or not $(a_i = 0)$. The aggregate mass of population that attacks is denoted by A such that

$$A = \int_0^1 a_i di.$$

[6] Also see Shadmehr and Bernhardt (2015), and Edmond and Lu (2017) on information manipulation in other political economy settings.

The strength of the status quo is represented by a random variable θ. The status quo survives if and only if $\theta \geq A$. The state is drawn from a commonly known probability distribution on $\Theta = [\underline{\theta}, \bar{\theta}] \subseteq \mathbb{R}$. The cumulative probability function (CDF) of the distribution $F(\cdot)$ is differentiable for every θ, and let $f(\theta)$ denote its density function.

If the agent does not attack, her payoff is zero. If the agent attacks, her payoff depends on her action and the regime status: she incurs positive cost $c \in (0, 1)$ regardless of the regime status, and if the regime is overthrown, she receives a benefit, which is normalized to be 1.[7] An agent's utility function is therefore

$$u(a_i, A, \theta) = a_i(\mathbb{1}\{\theta < A\} - c)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. We assume that

$$\bar{\theta} > 1 > \underline{\theta} \geq 0.$$

In other words, there are states ($\theta > 1$) in which the corresponding base game is dominance solvable with no attack. We assume $\underline{\theta} \geq 0$ to rule out the uninteresting case, but this assumption is not essential.[8]

Information structure. An information designer *commits* to disclosing information to the agents about the state θ. This is modeled as an information structure consisting of a signal space S and a state-dependent distribution over the signal profile $S^{[0,1]}$. For our purpose, it is sufficient to specify the resulting state-dependent distribution $\pi : \Theta \to \Delta(S)$ where $\pi(s|\theta)$ corresponds to the measure of agents receiving signal $s \in S$. Hence, the received signal s also corresponds to the agent's *type*. We assume the information structure is *anonymous*, i.e., agent i receives signal s in state θ with probability $\pi(s|\theta)$ for any i, s, and θ. Put differently, the designer is allowed to send differential signals to agents, but he is unable to discriminate agents based on their identities.[9] Without loss of generality, we further focus on the class of distributions where the density is almost everywhere well-defined and integrable, and thereby restrict our attention to policies under which the regime outcome is measurable in the information designer's information. In the rest of the paper, we use information structure and its resulting distribution of agents' types (S, π) interchangeably unless otherwise noted.

Bayesian game and solution concept. The combination of information structure and

[7] The benefit can be interpreted as ideology or pecuniary benefits that help to overcome the classic free-rider problem (Olson 1965). For example, agents may view their participation in an attack as beneficial for the society and therefore it directly adds to their utility. See chapter 2.3 of Acemoglu and Robinson (2005) for a comprehensive discussion.

[8] Alternatively, one can allow $\underline{\theta}$ to be negative, and when $\theta < 0$, the regime changes as long as one agent attacks. None of our analysis is affected.

[9] Our configuration of π, which determines with certainty the measure of agents receiving each signal, provides a tractable framework for the exposition of economics in our main results. Meanwhile, our analysis readily extends to the general and more complex cases without determinate measures or agent anonymity. We delegate the generalization to Section 4.

base game constitutes a Bayesian game, which proceeds as follows. First, θ is drawn by nature. Then, given an information structure indexed by (S, π), each agent i receives signal $s \in S$ according to π, and all agents simultaneously choose their actions. Agent i's *strategy* $a_i : S \to [0, 1]$ specifies the probability of attack. In a Bayesian Nash equilibrium, given a_{-i} and her own signal s, agent i attacks if and only if she strictly prefers to attack.

For a given information structure, there may be multiplicity due to the coordination nature of the base game. We solve for the information designer's *worst* Bayesian Nash equilibrium to capture the idea of adversarial/robust information design. That is, for each information structure, agents coordinate on a strategy profile such that the largest measure of agents attacks. In the remainder of the article, we refer to the information designer's worst Bayesian Nash equilibrium as *equilibrium*.

The information designer's problem is to choose (S, π) to induce an adversarial Bayesian Nash equilibrium which maximizes the regime's expected probability of survival.

3 Analysis

We begin with the equilibrium characterization for an arbitrary information structure.

Proposition 1. *For every (S, π), the induced Bayesian game has a unique equilibrium.*

We follow the familiar argument of iterated elimination of strictly dominated strategies (IESDS) to construct an equilibrium. Fix an information structure (S, π), we begin with the most aggressive strategy where all agents attack regardless of their signals. We identify a set of no-attack signals S_1 such that an individual agent finds attack to be dominated when receiving a signal in S_1. Then we examine an agent's incentive when she believes all other agents play a less aggressive strategy: attack if and only if their signals are outside of S_1. We identify another set of no-attack signals S_2 such that an agent finds it sub-optimal to attack when receiving signals in S_2. Since agents' actions are strategic complementary, the best response to a less aggressive strategy must be less aggressive, making $S_2 \supseteq S_1$. This iteration proceeds further for $S_3, S_4 \cdots$. As k goes to infinity, we obtain the maximal set of no-attack signals $S^* = \lim_{n \to \infty} S_n \subseteq S$. In doing so, we construct an equilibrium where an agent attacks if and only if his signal lies in $S \setminus S^*$.

The equilibrium probability of the status quo being overthrown is unique because we solve for the designer's worst equilibrium. But there may be a multiplicity in the set of non-attack signals. We further show there is a unique equilibrium. Note that one cannot apply the standard iterated dominance argument (Morris and Shin 1998) to prove the uniqueness. This is because we allow for an arbitrary information structure, and so there may not be signals serving as take-offs for the iterated domination for both actions. The uniqueness is indeed driven by our equilibrium selection rule. Suppose two distinct equilibria with different sets of no-attack signals S^* and S^{**}. Since two equilibria induce an identical probability of regime changes, both S^* and S^{**} must contain some exclusive signals respectively. We show that there must be another equilibrium where agents play weakly more aggressive than the following strategy: attack if and only if receiving signals from $S \setminus (S^* \cap S^{**})$. This is, once again, due to the strategic complementarity: a more aggressive strategy leads to a more aggressive best response. However, this equilibrium induces a strictly larger probability of regime change, which leads to a contradiction.

3.1 Main Result

Our main result is the characterization of the optimal information structure, which maximizes the probability of survival of the status quo. First, we introduce a class of information structures.

Definition 1. *An information structure (S, π) is a* ***local obfuscator*** *if*

1. *there is a non-empty subset of the state space $(\underline{\theta}', \bar{\theta}']$ being partitioned into a sequence of intervals $\{(\theta_{k+1}, \theta_k]\}_{k=0}^{\infty}$ where $\theta_0 = \bar{\theta}'$, and $\lim_{k\to\infty} \theta_k = \underline{\theta}'$,*
2. *the signal space S is such that $\{s_k\}_{k=1}^{\infty} \subset S$, and*
3. *the state-dependent distribution π is such that*

$$\begin{cases} \pi(s_1|\theta) = 1 & \text{if } \theta \in (\theta_1, \theta_0] \\ \pi(s_{k+1}|\theta) + \pi(s_k|\theta) = 1 & \text{if } \theta \in (\theta_{k+1}, \theta_k], \forall k > 1 \\ \pi(s_k|\theta) = 0, \forall k = 1, 2, ... & \text{if } \theta \notin (\underline{\theta}', \bar{\theta}'] \end{cases}.$$

In other words, if an information structure locally obfuscates agents, a set of adjacent states is categorized into a number of intervals, each of which corresponds to a unique signal. We interpret interval $(\theta_{k+1}, \theta_k]$ as the *face value* of signal s_{k+1}. When the state is $(\theta_{k+1}, \theta_k]$, an agent receives either signal s_{k+1} or a *slightly elevated* signal, s_k. Figure 1 visualizes an information structure that exhibits local obfuscation.

We refer to the obfuscation induced by the aforementioned information structure as *local* for two reasons. First, an agent can never distinguish states that belong to the same interval. Second, when an agent is misinformed, she receives an elevated signal corresponding to the interval just above the true one. The possibility of being locally obfuscated makes the agent skeptical. When receiving signal s_k, instead of taking the signal at face value, the agent believes that the true state is in either $(\theta_{k+1}, \theta_k]$ or $(\theta_k, \theta_{k-1}]$, and her posterior belief is derived by Bayes' rule. Moreover, the obfuscation creates information asymmetry and higher-order uncertainty among agents. An agent

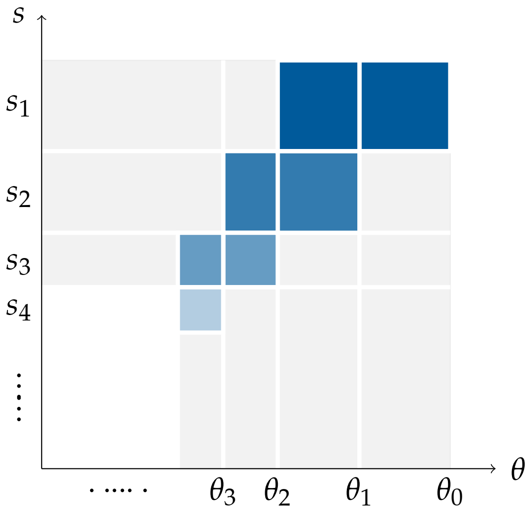

Figure 1: Illustration of local obfuscator. The horizontal axis represents states and the vertical axis represents signals and their face values. We use differential shades to distinguish information sets following different signals. When the state $\theta \in (\theta_1, \theta_0]$, every agent is truly informed by signal s_1 whose face value coincides with the interval containing the state. When $\theta \in (\theta_2, \theta_1]$, some agents receive signal s_2, but others receive signal elevated s_1. Similarly, when $\theta \in (\theta_{k+1}, \theta_k]$, some agents are receive signal s_{k+1}, but others receive the elevated signal s_k.

who receives signal s_k is uncertain not only about the interval that contains the true state but also about the signals received by other agents, making the coordination harder. The information designer can manage agents' posterior beliefs about other agents' signals, beliefs and therefore action profiles by manipulating the information structure.

We are ready to present our main result.

Theorem 1. *A local obfuscator (S, π^*) is an optimal information structure if*

1. *the sequence $\{\theta_k\}_{k=0}^{\infty}$ is such that $\theta_0 = \bar{\theta}$, $\theta_1 = 1$, θ_2 solves*

$$c \int_1^{\bar{\theta}} f(\theta) d\theta = (1-c) \int_{\theta_2}^{1} (1-\theta) f(\theta) d\theta, \tag{1}$$

and θ_k is recursively solved by

$$c \int_{\theta_{k-1}}^{\theta_{k-2}} \theta f(\theta) d\theta = (1-c) \int_{\theta_k}^{\theta_{k-1}} (1-\theta) f(\theta) d\theta, \tag{2}$$

for $k = 3, 4, \ldots$, and $\lim_{k\to\infty} \theta_k = \theta^*$ which solves

$$c \int_1^{\bar{\theta}} f(\theta) d\theta + \int_{\theta^*}^{1} (\theta - 1 + c) f(\theta) d\theta = 0, \tag{3}$$

2. *the signal space is given by $S = \{s_k\}_{k=1}^{\infty} \cup \{s_a\}$, and*

3. *the state-dependent distribution over S is such that*

$$\begin{cases} \pi^*(s_1|\theta) = 1 & \text{if } \theta \in (\theta_1, \theta_0] \\ \pi^*(s_{k+1}|\theta) = 1 - \pi^*(s_k|\theta) = \theta & \text{if } \theta \in (\theta_{k+1}, \theta_k] \cap \Theta, \forall k > 1 \\ \pi^*(s_a|\theta) = 1 & \text{if } \theta \in [\underline{\theta}, \theta^*] \cap \Theta \end{cases} .$$

Given (S, π^*), an agent attacks if and only if receiving signal s_a, and the status quo survives if and only if $\theta \in (\theta^*, \bar{\theta}] \cap \Theta$.

Theorem 1 says that there is an optimal information structure that exhibits local obfuscation. In other words, to maintain the status quo, the information designer needs only to *slightly exaggerate* the true state to *some* agents. The equilibrium regime status is fully determined by θ^*, which is pinned down by equation (3) whenever it has a solution. If equation (3) has no solution, the status quo survives for sure; otherwise, the regime status is state-dependent. When $\theta \leq \theta^*$, every agent receives signal s_a and attacks, and the status quo collapses. When $\theta > \theta^*$, agents are locally obfuscated and the status quo survives. Given the state $\theta \in (\theta_{k+1}, \theta_k]$ for some $k > 1$, the proportion of agents being deceived by an elevated signal s_k is $1 - \theta$, which decreases in θ. Hence, the information designer is more desperate to send elevated signals to more agents when the strength of the status quo becomes weaker. The optimal local obfuscator has a *global* impact. First, when $\theta > \theta^*$, it collapses all agents' attack by sending disinformation to a proportion of agents only. Second, it suppresses agents' attack in a large set of states through obfuscating nearby states.

For the sake of simplicity in notation, we will henceforth suppress the signal space S and refer to the optimal local obfuscator as π^*. The rest of this section is devoted to heuristically explaining the optimality of π^*.

The equilibrium outcome under π^*. We begin with an agent who receives signal s_1. Given her knowledge about π^*, she infers that the true state θ is either in $(\theta_1, \theta_0]$ or in $(\theta_2, \theta_1]$. If $\theta \in (\theta_1, \theta_0]$, the status quo survives regardless of the agents' coordinated action, making attack strictly sub-optimal. If $\theta \in (\theta_2, \theta_1]$, the regime changes only if a sufficiently large amount of agents attack. Since θ_2 balances equation (1), given s_1, the conditional expected benefit of attack does not exceed the cost even if all other agents attack for sure. Consequently, the agent does not attack.

Leveraging by s_1 being sent when $\theta > \theta_1$, the information designer creates a sequence of signals $s_2, s_3, ...$ that prevent agents from attacking in a large set of weaker states. To see it, we apply mathematical induction. Suppose that for every k, agent i does not attack when receiving signals $s_1, s_2, ..., s_{k-1}$. If agent i's signal is s_k, she believes that the state is either in $\theta \in (\theta_k, \theta_{k-1}]$ or in $\theta \in (\theta_{k+1}, \theta_k]$. In the former event, the measure of agents who also receive signal s_k is θ, and others receive signal s_{k-1}. By the induction hypothesis, agents who receive s_{k-1} do not attack. Thus, in this event, the status quo survives even if every agent who receives signal s_k attacks, making it strictly sub-optimal to attack. Moreover, by equation (2), the posterior belief that the true state is in $(\theta_k, \theta_{k-1}]$ is c; thus, even if every agent receiving s_k attacks, the probability of winning still cannot justify the cost of attack. Consequently, agent i does not attack as well.

As a result, if $\theta \in (\theta^*, \bar{\theta}]$, agents receive signals $s_1, s_2, ...$ and do not attack, and so the status quo survives. When they receive s_a, it is common knowledge that the state is in $[\underline{\theta}, \theta^*]$, so all agents attack, and the status quo is overthrown.

The Optimality of π^*. The optimal information structure indeed coincides with the one described in Figure 1. To explain why this is always the case, we have developed a "credit-discredit" system to describe the hierarchy of endogenously induced beliefs among the agents. In the process of IESDS that determines the agent equilibrium, whenever a state θ will certainly survive given the agents' current maximal rational coordination on attack, we say that θ can create credit by sending a self-identifying signal, in the sense of allowing for some weaker state θ' to survive by also sending that signal (possibly with a probability). Conversely, we say that θ' now creates discredit because, by mimicking θ, it weakens an agent's incentive to refrain from attacking the status quo.

We illustrate how the system works by a straightforward example. For any information structure, the credit created in the first round of IESDS is identical, which stems from a signal s sent by the invincible states $(1, \bar{\theta}]$. Now consider state $\theta = 0.9$, and suppose that it sends s to 20% of agents and a different signal s' to the remaining 80%. Clearly, the 20% will refrain from attack given s in fear of facing an invincible state, while the 80%, realizing this, will not attack given s' either because they expect a sure defeat. Hence $\theta = 0.9$ creates discredit by sending s, survives, and meanwhile creates credit by sending s'; some other states may then create discredit by sending s', survive, and create new credit in an analogous manner. Note that the standard obfuscation allocation — for instance in an information design using public announcements — terminates after the first round of IESDS, while our iterative construction may continue given a carefully designed information structure. This marks an important difference made by allowing discriminatory signals — the leveraged subset not only consumes credit but also creates credit for subsequent rounds of IESDS, and even more states may be saved along the process.

The credit-discredit system leads to an important property of the optimal information structure. Consider an arbitrary information structure and an arbitrary round of IESDS, at the beginning of which a certain amount of (net) credit, leftover from all previous rounds, is available at the information designer's disposal. The information designer's problem is then to select a subset of currently still vulnerable states to exploit the existing credit via creating discredit and thus survive, and at the same time create new credit on their own. Note that for any state with strength $\theta < 1$ to be in-

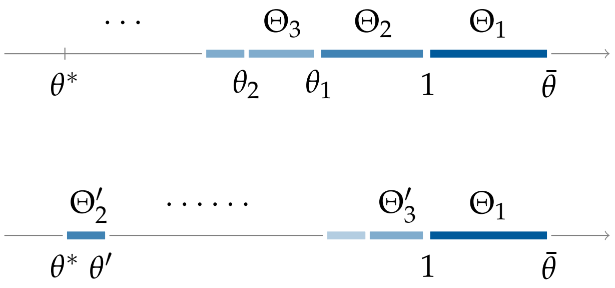

Figure 2: The upper panel corresponds to IESDS for the optimal local obfuscator. Denote $\Theta_1 = (1, \bar{\theta}]$ and Θ_k as the set of states being leveraged in the $k+1$th round of IESDS. As $k \to \infty$, every $> \theta^*$ state is leveraged. In the lower panel, the procedure is similar except in the first round. $\Theta'_2 = (\theta^*, \theta']$ where θ' is chosen to balance the credit constraint, $c \int_1^{\bar{\theta}} f(\theta)d\theta = (1-c)\int_{\theta_*}^{\theta'}(1-\theta)f(\theta)d\theta$. Obviously, the resulting measure of Θ'_2 is less than Θ_2. The choice of Θ'_2 further tightens the credit constraints in subsequent rounds, i.e. $\int_{\Theta'_k} \theta df(\theta) < \int_{\Theta_k} \theta df(\theta)$, making the measure of Θ'_{k+1} less than Θ_{k+1} for $k = 3, 4, 5...$

cluded in this round, it surely survives from attack as long as it sends a self-identifying signal (credit) to no more than a θ fraction of the agents, while the rest $1-\theta$ fraction of agents receive some signal mimicking a stronger state from the previous round (discredit). This argument reveals a nice duality: on the one hand, the maximum measure of credit it can offer is exactly θ, which is increasing in θ; on the other hand, the minimum measure of discredit it needs to create for survival is $1-\theta$, which is decreasing in θ. In other words, the information designer's conditional optimal choice — which maximizes the additional states that can be saved after this round — is to select the highest states possible. We thus obtain a recursive characterization of the unconditional optimum, summarized by (1) and (2) where the right-hand sides of equations correspond to the "credit production" in the kth round, while the left-hand sides correspond to the "credit consumption." It then implies that at optimum the surviving states form one unique interval $(\theta^*, \bar{\theta}]$. Figure 2 demonstrates the optimality to leverage states monotonically in IESDS.

By this property, we thus obtain an explicit upper bound for the status quo's survival probability, which also identifies a lower bound for a surviving state at optimum, by a straightforward necessary condition which leads to (3) at optimum:

$$\frac{\int_1^{\bar{\theta}} f(\theta)d\theta + \int_{\theta^*}^1 \theta f(\theta)d\theta}{\int_{\theta^*}^1 (1-\theta)f(\theta)d\theta} \geq \frac{1-c}{c}.$$

This inequality implies that the ratio between the total measures of credit and discredit created by $(\theta^*, \bar{\theta}]$ must be at least $\frac{1-c}{c}$. Intuitively, a sufficiently large proportion of credit in the set of no-attack signals need to be provided, to hold an agent at least indifferent between attacking and not.

Finally, we verify that π^* achieves exactly the maximum survival probability of the status quo by direct calculation. This can be easily seen by summing up (1) and (2) over k to arrive at (3) at the limit. In π^*, the ratio between credit and discredit in every round of IESDS is kept at precisely $\frac{1-c}{c}$, which automatically preserves the same ratio between the total measures.

The necessity of multiple signals. Although the optimal information structure essentially produces a set of attack signals and another set of no-attack signals, the maximum survival probability of the status quo cannot be reached by pooling all signals into binary recommendation signals. To see the logic, first notice that the classic revelation principle/Bayes-correlated-equilibrium approach (See Bergemann and Morris (2016) and Taneva (2019)) does not apply if one focuses on the information designer's worst equilibrium. More importantly, the multiple (and possibly infinite) rounds of IESDS are necessary to maximize the status quo's survival probability. The binary recommendation is at best equivalent to the first round of IESDS under the optimal information structure.

3.2 Comparative Statics

In this section, we take a closer look at the optimal local obfuscator π^*. Recall that in the construction of an optimal local obfuscator, the subset of invincible states $(1, \bar{\theta}]$ is critical. When the state lies in this interval, no attack is a strictly dominant strategy. The local domination in this subset serves as a take-off to construct a sequence of signals extending the suppression of agents' coordination to lower states. This ripple effect leveraged by invincible states naturally depends on two primitives, the cost of attack, c and the likelihood that attack being dominated, $F(1)$. In this section, we conduct comparative static analysis on these two primitives and examine the advantage of local obfuscation relative to public propaganda.

Public Signals. As a benchmark, we first derive the optimal public information structure, i.e., for every state θ, signals received by any two agents i, j must be identical. Straightforwardly, it is optimal to set the signal space to be binary, $S = \{s_a, s_n\}$, and broadcast an attack signal s_a if $\theta \leq \theta^\dagger$ and a no-attack signal s_n otherwise for some cutoff $\theta^\dagger$ solving

$$c = \frac{F(1) - F(\theta^\dagger)}{1 - F(\theta^\dagger)}. \tag{4}$$

The right-hand side of equation (4) is an agent's expected benefit if she attacks given that $\theta > \theta^\dagger$ and all other agents attack. Given the no-attack signal s_n, the agent believes that $\theta > \theta^\dagger$, and finds not to attack to be weakly dominant. This is because when $\theta \in (1, \bar{\theta}]$, attack is a strictly dominated strategy. Obfuscating states on $(\theta^\dagger, \bar{\theta}]$ makes attack an unwise choice given s_n.

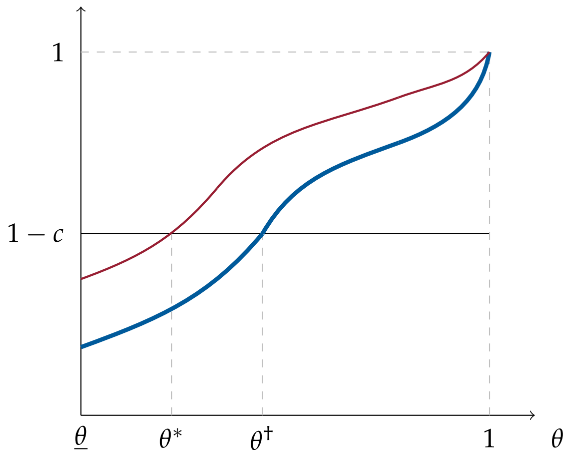

Figure 3: The comparative statics. The thick blue curve represents the right-hand side of equation (5), and the thin red curve represents the right-hand side of equation (6). When c increases, the black curve $1-c$ is shifted down for every θ, and so both θ^* and $\theta^\dagger$ decrease.

To ease the discussion of comparative statics, we rewrite equation (4) as

$$1-c = \frac{1-F(1)}{1-F(\theta^\dagger)}, \tag{5}$$

which is plotted in Figure 3. Naturally, the cutoff value $\theta^\dagger$ is decreasing in c. When $c \geq F(1)$, we have $\theta^\dagger = 0$: agents never attack, and the status quo always survives. When the cost of attack falls, the coordination becomes easier, and the status quo survives in a smaller set of states. As $c \to 0$, $\theta^\dagger \to 1$, and the status quo fails whenever $\theta \notin (1, \bar{\theta}]$. In this case, the leverage caused by the local domination in $(1, \bar{\theta}]$ on lower states vanishes. We can also increase the total measure of the local domination interval $(1, \bar{\theta}]$, and decrease the probability density for each $\theta \leq 1$ in an arbitrary way to balance the total probability to be one. By equation (5), the probability that the status quo survives $1 - F(\theta^\dagger)$ changes proportionally. That is, raising the measure of the local dominance interval has a constant multiplier effect on maintaining the status quo. Intuitively, this multiplier effect is stronger when the cost of attack is larger. We summarize the comparative statics results in the following proposition.

Proposition 2.A. *In an optimal public information structure, the ex ante probability that the status quo survives, $1 - F(\theta^\dagger)$ has the following properties.*

1. *It increases in c, converges to $1 - F(1)$ as $c \to 0$, and equals one if $c \geq F(1)$.*

2. *When $1 - F(1)$ increases, and $f(\cdot)$ decreases arbitrarily and accordingly for $\theta < 1$, $1 - F(\theta^\dagger)$ increases at a constant rate $\frac{1}{1-c}$. When $1 - F(1) \to 0$ and $f(\cdot)$ increases arbitrarily and accordingly for $\theta < 1$, $1 - F(\theta^\dagger)$ converges to 0.*

It is worth noting that the second statement immediately implies that the status quo's survival probability increases in F in the sense of first-order stochastic dominance, i.e. if the distribution of θ becomes G which first-order stochastic dominates F, the status quo survives with a higher probability under an optimal public information structure.

Local Obfuscation. Now we turn to the comparative statics on optimal local obfuscation. Rewrite equation (3) as

$$1 - c = \frac{1 - F(1)}{1 - F(\theta^*)} + \frac{\int_{\theta^*}^{1} \theta f(\theta) d\theta}{1 - F(\theta^*)}. \tag{6}$$

Compared to equation (5), equation (6) has a new term on the right-hand side. It captures the total benefit of using local obfuscation through a sequence of signals. Notice that it is simply, $\mathbb{E}(\theta | \theta \in (\theta^*, 1])$ the conditional expectation $\theta \in (\theta^*, 1]$, the states being leveraged by the local dominance interval $(1, \bar{\theta}]$.

Once again, the cutoff value θ^* is depicted in Figure 3. Higher cost of attack makes the coordination more difficult, and therefore lowers the cutoff state θ^*. Hence, θ^* decreases in c, and converges to 1 as $c \to 0$. If

$$c \geq F(1) - \int_{\underline{\theta}}^{1} \theta f(\theta) d\theta, \tag{7}$$

the agents never attack and the status quo never fails. Notice that in this case, the ex ante optimal local obfuscator is also *ex post optimal* to the designer, so it remains credible even if the designer has no commitment power.

Equation (6) also suggests that when the dominance interval $1 - F(1)$ increases, the rate of change of $1 - F(\theta^*)$ is no longer constant; in fact, if we allow arbitrary ways of decreasing the respective probability density for $\theta \leq 1$, the change is even not necessarily positive. This is because when the change has an ambiguous effect on both θ^* and $\mathbb{E}(\theta | \theta \in (\theta^*, 1])$. Nevertheless, the monotonicity of survival probability under first-order stochastic dominance is preserved. Indeed, when the state distribution becomes more skewed towards stronger states, more credit and less discredit are created for every given measure of surviving < 1 states. Thus the information designer may prevent more states from being attacked by enrolling them into the iterated process.

Under an optimal information structure $1 - F(\theta^*)$ is bounded away from 0 even if the dominance interval converges to measure 0. The intuition is that a non-public information structure can leverage much more states — those in the dominance interval, as well as those that survive in the subsequent rounds of IESDS. Note that the states below but sufficiently close to 1 actually produce more leverage for subsequent states than consumed from a previous round of IESDS to save them: in particular, every state θ satisfying $\theta > 1 - c$ lies in this category. Then no matter how small $1 - F(1)$ is, it will start the iterated reasoning process that keeps saving lower states, and the process will never stop before $\theta < 1 - c$. Therefore $1 - c$ presents an explicit upper bound for

θ^*, meaning that as long as $\theta \in [1-c, 1]$ with a significant probability, the status quo survives also with a significant probability however small the measure of invincible states is.

The comparative statics is summarized as follows.

Proposition 2.B. *Under an optimal information structure, the ex ante probability that the status quo survives, $1 - F(\theta^*)$ has the following properties.*

1. *It increases in c, converges to $1 - F(1)$ as $c \to 0$, and equals one if $c \geq c^*$.*
2. *Suppose that G first-order stochastically dominates F, and let θ^{**} denote the lower bound of surviving states under the corresponding optimal local obfuscator given G. We have $1 - G(\theta^{**}) \geq 1 - F(\theta^*)$.*
3. *Consider $\{F_n\}_{n \in \mathbb{N}^+}$ (with f_n and θ_n^* defined correspondingly) such that $\lim_{n\to\infty} 1 - F_n(1) = 0$, and suppose that $\liminf_{n\to\infty} f_n(\theta) > 0$ for all $\theta \in \hat{\Theta}$, for some non-empty $\hat{\Theta} \subset [1-c, 1]$. Then $\liminf_{n\to\infty} 1 - F_n(\theta_n^*) > 0$.*

Public vs Private Signals. We are now ready to discuss the advantage of the local obfuscation compared to the public signal (propaganda). One way to examine the advantage is to look at $F(\theta^\dagger) - F(\theta^*)$, the measure of the set of states that coordination is crushed under local obfuscation only.

Proposition 3. *The advantage of local obfuscation relative to public propaganda $F(\theta^\dagger) - F(\theta^*)$ has the following properties:*

1. *It is non-negative for every c, and strictly positive when $c < F(1)$.*
2. *It is increasing in c.*
3. *Suppose that $\{F_n\}_{n \in \mathbb{N}^+}$ (with f_n, $\theta_n^\dagger$ and θ_n^* defined correspondingly) satisfy the conditions in Proposition 2.B. Then $\liminf_{n\to\infty} F_n(\theta_n^\dagger) - F_n(\theta_n^*) > 0$.*

Under the optimal public information structure, even fewer states survive than under π^* after the first round of IESDS. The reason is that the public information structure inevitably wastes some credit provided by $(1, \bar{\theta}]$. For the sake of argument, consider a hypothetical measure 1 of some state $\theta < 1$. The public information structure can save θ from a regime change only by designing for it the same signal as some > 1 state, therefore inducing *all* agents to refrain from attacking. In other words, θ creates discredit of measure 1 as well. Under π^*, however, θ only mimics some > 1 state towards $1 - \theta$ fraction of the agents, reducing the measure of discredit produced to only $1 - \theta$. The remaining measure of θ then leaves room for more < 1 states to fill with their discredit and survive. Hence as long as the optimal public information structure saves a proportion of states < 1, π^* must be strictly preferred by the information designer

(Property 1). It then follows directly from this argument that the additional survival probability induced by π^* over the optimal public information structure in the first round of IESDS, as well as that in every subsequent round under π^*, is increasing in c, which leads to Property 2. Note also that both $\theta^\dagger$ and θ^* approach 1 as $c \to 0$; that is, even when non-public information structures are available, an infinitesimal cost always renders information design futile.

Property 3 highlights a significant difference between public and non-public information structures in an extreme scenario. Although $F(\theta^\dagger) - F(\theta^*)$ may not be monotone in $1 - F(1)$, the probability measure of invincible states, it does remain bounded away from 0 as the measure gradually becomes negligible. This result implies that using non-public signals indeed bears a unique advantage, which does not vanish even when the optimal public signal becomes almost ineffective. However small the measure of invincible states is, it creates a significant ripple effect by the infinite rounds of IESDS under π^*. The starkest contrast arises when $c > 1 - \int_{\underline{\theta}}^1 \theta f(\theta) d\theta$ and $1 - F(1) \to 0$: almost no state survives under the optimal public information structure, but all states survive under optimal local obfuscation!

Another way to understand the advantage of local obfuscation is through the different necessary levels of attacking cost to save the same set of states under different information structures. Specifically, we can define

$$c^\dagger(\theta) \equiv \frac{F(1) - F(\theta)}{1 - F(\theta)}$$
$$c^*(\theta) \equiv \frac{F(1) - F(\theta)}{1 - F(\theta)} - \frac{\int_\theta^1 \theta f(\theta) d\theta}{1 - F(\theta)}$$

as the corresponding cost thresholds. They represent the lowest cost under which states $[\theta, \bar{\theta}]$ survive given the corresponding information structure. Then we measure the advantage of local obfuscation by $c^\dagger(\theta) - c^*(\theta) = \mathbb{E}(\theta' | \theta' \in [\theta, 1]) > 0$. This equality means that when the information designer switches from the optimal public information structure to the optimal non-public one while holding the status quo's survival probability constant, each agent's attacking cost can at most be allowed to decrease by $\mathbb{E}(\theta' | \theta' \in [\theta, 1])$, the expected strength of the states being "saved" (as compared to the > 1 states that always survive).

To interpret the equality, first note that states $[\theta, 1)$ only produce discredit under the optimal public information structure, which accounts for $\frac{F(1)-F(\theta)}{1-F(\theta)}$; the difference between the cost thresholds thus can be regarded as capturing additional credit created in the iterated process under π^*. One may then examine the condition required for a state with this expected strength, $\tilde{\theta} = \mathbb{E}(\theta' | \theta' \in [\theta, 1])$, to survive under π^* after its round in the iterated reasoning process. Namely, the ratio of credit created in its signal must be at least $1 - c$. We already know that this ratio coincides with its strength $\tilde{\theta}$; therefore an increase in $\tilde{\theta}$ allows for a decrease in c, in terms of saving an identical set of states, by exactly the same amount.

4 On Optimal Local Obfuscation

Now we turn to the robustness of using the optimal local obfuscator. First, we show that the unique optimal information structure exhibits local obfuscation when the signal space is restricted to be finite. Second, we discuss that the optimal local obfuscator remains optimal even if the designer is allowed to target agents according to their identities. Finally, we argue that a local obfuscator is robust to perturbation, such as private communication and private signal of agents, which undermines the full control of the information structure.

4.1 Restricting Complexity

Theorem 1 not only identifies the unique lower bound of the state where the status quo can survive at optimum, θ^*, but also provides an intuitive implementation even if no prior knowledge about θ^* is available. But it is worth noting that the optimal local obfuscator is not the *only* information structure securing the status quo's survival for $\theta > \theta^*$.

To understand the multiplicity of optimum, recall the "credit-discredit" interpretation. The optimal local obfuscator not only maximizes the credit production in every round of IESDS, but also uses stocking credit most economically, i.e. saves the most states given the credit constraint in each round. Nevertheless, alternative designs may exist under which the same *overall* amounts of credit and discredit are created as under the optimal local obfuscator, but different amounts occur in *each round* of IESDS. In such a design, the probability of the status quo's survival after the first k rounds of IESDS is strictly smaller than in π^* regardless of k; only as the process of IESDS takes infinitely many rounds and the marginal production of credit diminishes to zero, the gap becomes negligible as the procedure forwards.

We give a numerical example below, where $c = 1/6$, and θ is uniformly distributed on $\Theta = [0, 1.1]$. We consider a design that differs from π^* in that it identifies the newly surviving states in the second round of IESDS from θ^* upwards instead of from those in the first round downwards. The result is depicted in Figure 4: after the deviation in the second round, the probability of status quo's survival under π' is always strictly smaller than under π^* for any k, but will converge to the same limit as $k \to \infty$.

The above analysis, together with the intuition of Theorem 1, leads to the next proposition. The optimal local obfuscator π^* contains infinitely many signals $\{s_k\}$ preventing the status quo from attack, making the design very complex. In practice, it is natural to believe that the information designer is constrained to using *finite signals*. Given the restriction on the complexity of the signal space, we show that an optimal

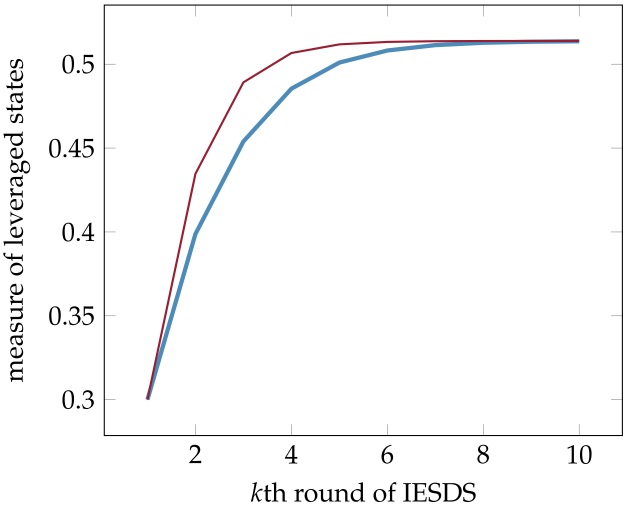

Figure 4: The horizontal axis represents the round of IESDS, and the vertical axis represents the cumulative measure of states being leveraged until each round. The thin red curve corresponds to the optimal local obfuscator π^*, while the thick blue curve corresponds to the alternative information structure π'. The total measure of states being leveraged under π' falls behind that under π^* since the second round of IESDS, but it eventually starts to catch up. When $k = 8$, the difference already shrinks to 0.0003.

information structure must exhibit local obfuscation.

Proposition 4. *For $n = 2, 3, \cdots$, let π_n denote the following state-dependent signal distribution:*

$$\begin{cases} \pi_n(s_1|\theta) = 1 & \textit{if } \theta \in (\theta_1, \theta_0] \\ \pi_n(s_k|\theta) = 1 - \pi_n(s_{k-1}|\theta) = \theta & \textit{if } \theta \in (\theta_k, \theta_{k-1}] \cap \Theta, \forall k = 2, \cdots, n-1 \\ \pi_n(s_a|\theta) = 1 - \pi_n(s_{n-1}|\theta) = \theta & \textit{if } \theta \in (\theta_n, \theta_{n-1}] \cap \Theta \\ \pi_n(s_a|\theta) = 1 & \textit{if } \theta \in [\underline{\theta}, \theta_n] \cap \Theta \end{cases}.$$

where $S = \{s_k\}_{k=1}^{n-1} \cup \{s_a\}$. Suppose that the information designer is restricted to using S that contains at most n elements; then either

1. *π_n is the unique optimal information policy, or*
2. *under an optimal information policy, no agent ever attacks and the status quo always survives.*

The argument underlying Proposition 4 is centered on maximizing the ripple effect created by the initial credit from $\theta \in (1, \bar{\theta}]$. When only finite signals are available, the agents only go through finite rounds of IESDS. In terms of credit creation, the iterated reasoning process among agents resembles money creation in the banking system to a certain extent. Intuitively, a certain amount of credit created in an earlier round proves more "useful" to the information designer than the same amount of credit in a later round, because it generates a larger sum of additional credit through the remaining rounds. By induction, the optimal information structure must seek to create maximum possible credit in each round sequentially, which uniquely corresponds to π_n.

4.2 Relaxing Determinate Measures and Anonymity

Our main results extend to the following general environment. Let S be a compact metric space; the information designer's proposed information structure π is now a mapping from Θ to $\Delta(M(S))$, where $M(S) \subset \{S^{[0,1]}\}$ is a set of integrable functions with codomain S. This configuration allows for (1) arbitrary correlation — given measurability across states, agents and signals — among signals and (2) information structures that target particular agent groups.

With a slight abuse of notation, we adopt the notation π^* for the local obfuscator specified in Theorem 1. We show that its optimality is preserved.

Corollary 1. *π^* remains optimal in the above environment.*

We leave the proof of this result to the Online Appendix. The main intuition is based upon an alternative characterization of the information designer's problem. Since the information designer aims to maximize the ex-ante probability of the status quo's survival, i.e. the probability measure of the surviving states, we can without loss of generality re-label each state θ as a multiple replica of itself bearing a total density of $f(\theta)$, each representing the same state under a realized measure distribution of signals. Given such a distribution, the state either survives or falls with certainty, in which case we can readily apply the proof of Theorem 1. On a more abstract level, it is only reasonable that as agents are coordinating on the information designer's least preferred equilibrium, introducing no correlation among signals will never hurt. Therefore, compared to the simplest i.i.d. information structure, the ability to target specific agents or to create arbitrary correlation yields no extra leverage for the information designer.

4.3 Exogenous Private Signals

A key distinction between our work and the literature of information design in coordination games is that the latter often assumes a structure of exogenous private signals among agents, so that heterogeneous private beliefs exist even without the designer's input. In our paper, the designer is interpreted as an informational autocratic, and agents are interpreted as citizens. In modern age, it is unrealistic to think that citizens have zero access to alternative news sources. It is therefore reasonable to consider the robustness of the optimal local obfuscator when agents receive exogenous private signals.

Needless to say, when agents' private signals are sufficiently informative, the designer's information manipulation will become fruitless. Here we provide a sketched argument that introducing not-very-informative exogenous private signals does not remove the optimality of local obfuscation. A general and formal analysis requires committing to specific a private information structure, which is left for future research.

Informed Elites. One of the simplest but meaningful ways of imposing exogenous private signals is to assume that a fraction $\Delta > 0$ of agents, either randomly or deterministically selected, knows the true state with certainty. In applications, these truth-knowing agents can be regarded as a group of "informed elite" as in Guriev and Treisman (2019). Our Theorem 1 can be directly applied to characterize the optimal information structure, with the minor alteration that every < 1 state above θ^* (which is endogenously determined) sends an elevated signal with probability $\frac{1-\theta}{1-\Delta}$ instead of $1-\theta$. It is easy to verify that θ^* is increasing in Δ; thus a large number of informed elites is not a good news to the informational autocratic.

General Specifications. An environment with more general exogenous private signals bears similar essence in logic. With potentially heterogeneous private signals, a fraction of agents will have better or more optimistic (in the sense that θ is more likely to be low) information about θ and thus become harder to discourage from attacking when $\theta < 1$. Our iterated credit-discredit system remains valid, but the designer has to deliberately shrink the range of θ in each round of IESDS and at the same time make those θ send an elevated signal more often, to once again guarantee that there is never a sufficient measure of agents who may coordinate on attacking. Of course, additional complication arises when the distribution of exogenous signals imposes implicit and non-standard constraints on credit creation in the iterated process, which may render the characterization of optimum less tractable. See Inostroza and Pavan (2018) for a discussion on information design with normally distributed exogenous signals.

4.4 Private Communication

Our final remark regards the robustness of local obfuscation when the designer cannot fully control the information structure. In the multi-agent information design settings, it is well known that using private signals can strictly improve the persuasion outcome as discussed in the introduction. It is often criticized that private communication of agents makes it impossible for the designer to perfectly differentiate agents' information in a significant amount. It is worth pointing out that our information structure is robust to private communication: collapsing local obfuscation requires a large proportion of agents to exchange a substantial amount of information to resolve both fundamental and strategic uncertainty.

To fix the idea, we use a stylized model extension to heuristically illustrate how the main economics of local obfuscation is preserved under limited private communication. The formal analysis is similar to the baseline model, so it is omitted. Suppose that, after receiving the signal sent by the designer, each agent is randomly paired with another agent, with whom she shares her signal. Under local obfuscation, this implies that an agent will end up in one of three possible information sets: two identical low signals that induce a optimistic belief; two identical high signals that induce a pessimistic belief; or two different signals that reveal the true state. To maintain the iterated reasoning process produced by local obfuscation as before, the designer only needs to appropriately increase the probability of sending an elevated signal by each < 1 state – and decrease the range of states that do so – in every round of IESDS, so that enough pessimistic agents will refrain from attacking. In this way, even a truth-knowing agent will not attack because she realizes that the fraction of peers that can possibly coordinate is never sufficient.

The optimal information structure in this setting remains an open question, as communication makes it potentially worthwhile for some state to send more than two signals. However, on the one hand, our above argument suggests that if the probability of each agent meeting another is arbitrarily small, the designer can always use an adjusted local obfuscator to reach a survival probability arbitrarily close to the one at optimum without communication. On the other hand, if each agent can share her information with more and more peers, it becomes harder and harder for a < 1 state to create credit through an iterated reasoning process. At the limit, if an agent meets a positive measure of other agents, common knowledge on the signal distribution arises. The optimal information structure then coincides with the optimal public propaganda.

To summarize, as long as the private communication among agents is limited, a local obfuscator is virtually optimal. In our opinion, the private information exchange is insufficient to overturn local obfuscation in many political economy settings. The reasons are twofold.

Indistinguishable nearby states. When the obfuscation is local, only signals representing nearby states are sent simultaneously. It is natural to believe that distinguishing nearby states is more difficult/costly to the agent than distant states. (See Hébert and Woodford (2017), Pomatto et al. (2018), Morris and Yang (2019), and Guo and Shmaya (2019) for formal discussion.) Thus, a slightly exaggerated signal is unlikely to be detected even if agents are allowed to privately verify the signals through a communication or private signals. To fully address this issue is beyond the scope of this paper; we therefore leave it for future research.

Echo Chamber. Second, people are more likely to communicate with those they interact with, which in turn creates "echo chambers" that prevent people from being exposed to information that contradicts their preexisting beliefs (see Levy and Razin (2018), Lipnowski and Sadler (2019), and Li and Tan (2019)). In our model, imagine that agents are divided into several chambers, and information exchange is allowed only within a chamber.[10] Since agents from the same chamber tend to share polit-

[10] In practice, information exchange between people with different political views is rare, hard to

ical views and be exposed to similar information sources, it is realistic to allow the designer to send target signals based on agents' chambers. Some randomly selected chambers receive the true signal and others receive the elevated signal. The designer is restricted to sending identical signal to agents from the same chamber, and so the maximum survival probability of the regime cannot be achieved, but the structure of local obfuscation remains.

5 Conclusion

Our analysis has shown that when the information designer has extensive power in information design, in particular when it can endogenously determine the structure of noise in the agents' information, the optimal persuasion scheme takes a simple and intuitive form. The information designer randomizes between honesty and deceit, which takes the particular form of local obfuscation. We believe that our stylized framework can be enriched to build a research agenda on many related topics, including competitive information designers, dynamic persuasion and communication among agents.

A Appendix: Proofs of Main Results

A.1 Proof of Proposition 1

We prove the proposition through a number of Lemmas. We begin with an order on the strategy space.

Definition 2. *For i's two strategies $\mathbf{a}_i$ and $\mathbf{a}'_i$, we denote that $\mathbf{a}_i \geq \mathbf{a}'_i$ if $a_i(s) \geq a'_i(s)$ for every $s \in S$, and that $\mathbf{a}_i > \mathbf{a}'_i$ if $a_i(s) \geq a'_i(s)$ for every $s \in S$ and $a_i(s) > a'_i(s)$ for some $s \in S$. We say $\mathbf{a}_i$ is **more aggressive** than $\mathbf{a}'_i$.*

The following Lemma regards i's best response given s. It is an immediate consequence of strategic complementarity among agents' actions. It says that when every other agents' strategies become more aggressive, an agent's best response is either unchanged or more aggressive.

Lemma 1. *Consider two strategy profiles of agents other than i, $\mathbf{a}_{-i}$ and $\mathbf{a}'_{-i}$. Suppose that $\mathbf{a}_j \geq \mathbf{a}'_j$ for every $j \neq i$, and that $\mathbf{a}_j > \mathbf{a}'_j$ for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. If it is optimal for agent i to attack given $s \in S$ and $\mathbf{a}'_{-i}$, it is also optimal to attack given s and $\mathbf{a}_{-i}$. Similarly, if it is optimal for agent i not to attack given s and $\mathbf{a}_{-i}$, it is also optimal not to attack given s and $\mathbf{a}'_{-i}$.*

motivate, and sometimes counter-productive. See Bail et al. (2018) for a field experiment that shows exposing people to opposing views on social media only increases political polarization.

Proof. We prove the first part of the lemma. The proof of the second part is almost identical and therefore omitted. Fix $s \in S$, the signal of agent i. Suppose that it is optimal for agent 1 to attack given $\mathbf{a}'_{-i}$ and signal s, and suppose that $\mathbf{a}_j \geq \mathbf{a}'_j$ for every $j \neq i$, and that $\mathbf{a}_j > \mathbf{a}'_j$ for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. We must have

$$\begin{aligned} c &< \int_{\Theta} \left(\frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]\setminus\{i\}} \int_S a'_j(v)\pi(v|\theta)dvdj\} \right) d\theta \\ &\leq \int_{\Theta} \left(\frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]\setminus\{i\}} \int_S a_j(v)\pi(v|\theta)dvdj\} \right) d\theta \end{aligned}$$

where the first inequality holds because of the optimality of attack given s and $\mathbf{a}'_{-i}$, and the second inequality holds because $\mathbf{a}_j \geq \mathbf{a}'_j$ for every $j \neq i$, and that $\mathbf{a}_j > \mathbf{a}'_j$ for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. Thus, agent i finds it optimal to attack given signal s and $\mathbf{a}_{-i}$. □

Now we are ready to address the equilibrium existence.

Lemma 2. *For any (S, π), there exists an equilibrium.*

Proof. Fix (S, π), we construct an equilibrium through *iterated elimination of strictly dominated strategies (IESDS)*. We begin with the strategy profile that everyone attacks, denoted as $a_i^0(s) \equiv 1$ for every i, and $s \in S$. For every i, define $S_1 \subseteq S$ as the set of signal s such that

$$\int_{\Theta} \frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]} \int_S a_j^0(v)\pi(v|\theta)dvdj\} d\theta \leq c. \tag{8}$$

That is, if agent i receives signal $s \in S_1$, he weakly prefers not to attack even if all other agents attack for certain. Define

$$a_i^1(s) = \begin{cases} 0 & \text{if } s \in S_1 \\ 1 & \text{otherwise.} \end{cases}$$

which is less aggressive than $\mathbf{a}_i^0$. By Lemma 1, an agent i weakly prefers not to attack if all other agents attack if and only if they receive signals in S_1.

For $k = 2, 3, \cdots$, define $S_k \subseteq S$ as the set of signal s such that

$$\int_{\Theta} \frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]} \int_S a_j^{k-1}(v)\pi(v|\theta)dvdj\} d\theta \leq c.$$

and define

$$a_i^k(s) = \begin{cases} 0 & \text{if } s \in S_k \\ 1 & \text{otherwise.} \end{cases}$$

Notice that $\mathbf{a}_i^k$ becomes less aggressive as k increases. By Lemma 1, $S \supseteq S_k \supseteq S_{k-1}$ for every k. At the limit as $k \to \infty$, the set $S^* = \lim_{k\to\infty} S_k$ exists, and $S^* \subseteq S$. Also, define:

$$a_i^*(s) = \begin{cases} 0 & \text{if } s \in S^* \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

for each agent i. Notice that for an arbitrary (S, π), S_1 may be empty. In that case, $S_k, S^* = \emptyset$.

Next we show that the strategy profile specified in (9) is indeed an equilibrium. First, by the construction of S^*, an agent prefers attacking when receiving every signal in $S \backslash S^*$ given other agents follow the strategy specified in (9). Second, we show that for any non-empty S^*, given that the other agents follow the strategy in (9), an individual agent i strictly prefers not attacking for every signal in S^*. The proof is straightforward. Pick any $s \in S^*$, there exists a unique k such that $s \in S_k \backslash S_{k-1}$. By the definition of S_k, given that the other agents $j \neq i$ follow $\mathbf{a}_j^{k-1}$ and do not attack if and only if receiving signals in S_{k-1}, an individual agent prefers not attacking when receiving signals in $S_k \backslash \bar{S}_{k-1}$. Then by Lemma 1, if other agents $j \neq i$ follow a less aggressive strategy $\mathbf{a}_j^* < \mathbf{a}_j^{k-1}$ and do not attack if and only if receiving signals in $S^* \supseteq S_{k-1}$, an agent must prefer not attacking when receiving signals in $S_k \backslash S_{k-1}$. Thus, for every $s \in S^*$, $a_i^*(s) = 0$. Hence, we have the desired result. □

The definitions above guarantee a unique series of $\{S_k\}$ and a unique S^*. In what follows, we show that $a_i^*(s)$ is the unique equilibrium as well.

Lemma 3. *For any (S, π), there is a unique equilibrium.*

Proof. We first introduce a useful function form representing the agent's strategy. Define the expected measure of agents who attack in state θ as

$$A_\theta = \int_S a_i(v)\pi(v|\theta)ds, \forall \theta \in \Theta.$$

Pick any two agents i, j, in every equilibrium, the expected payoff of attacking when receiving signal s is

$$\int_\Theta \frac{f(\theta)\pi(s|\theta)}{\int_\Theta f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < A_\theta\} d\theta - c,$$

which is invariant to the identity of the agent. Thus, every equilibrium strategy must be *symmetric*, i.e. for every s, $a_i(s) = a_j(s) = a(s)$, which takes value either 0 or 1. For the sake of contradiction, suppose that for (S, π), there are two distinct equilibria a, a'. Let $\{s|a(s) = 0\}$ denote the set of signals the agents do not attack in equilibrium a and $\{s|a'(s) = 0\}$ denote the set of signals the agents do not attack in equilibrium a'. By the hypothesis that a and a' are distinct equilibria, $\{s|a(s) = 0\} \neq \{s|a'(s) = 0\}$.

Moreover, they induce an identical probability of a regime being overthrown, so each set must contain some exclusive signals. Consider the following strategy a'' defined as follows:

$$a''(s) = \begin{cases} 0 & \text{if } s \in \{s|a(s) = 0\} \cap \{s|a'(s) = 0\} \\ 1 & \text{otherwise} \end{cases}$$

which is strictly more aggressive than a and a'. By Lemma 1, an individual agent i receiving a signal in $S \backslash (\{s|a(s) = 0\} \cap \{s|a'(s) = 0\})$ prefers attacking if every other agent is adopting strategy a''. Note that an equilibrium always exists, thus there must exist an equilibrium where the agents play at least as aggressively as a''. In such a case the regime changes with a greater probability than both in a and in a', which is a contradiction. □

The combination of Lemmas 1-3 yields Proposition 1.

A.2 Proof of Theorem 1

Step 1. *We define two series which will be useful in the following analysis. Given any information policy, these series are identified through IESDS; and they characterize the agents' iterative reasoning in coordination. Series $\{S_k\}_{k=1}^{\infty}$ is drawn from the proof of Proposition 1; it contains the signal sets which the agents refrain from attacking after the kth round of IESDS. Series $\{T_k\}_{k=1}^{\infty}$ satisfies the following condition: $\cup_{n=1}^{k} T_n$ contains the states that survive before the kth round of IESDS.*

Define state set $T_1 = (1, \bar{\theta}]$. By the definition of S_1, for every $s \in S_1$, s induces the following posterior: the probability that the true state is in T_1 is larger than $1 - c$, i.e.

$$\Pr(\theta \in T_1 | s) \geq 1 - c, \forall s \in S_1.$$

Next, we recursively define T_k as the set of states θ where more than $1 - \theta$ measure of agents receive signals in S_{k-1} , i.e.

$$T_k \equiv \{\theta \in \Theta : \int_{s \in S_{k-1}} \pi(s|\theta) ds > 1 - \theta\}$$

for every $k = 2, 3, ...$ Then, by the definition of S_k, for every $s \in S_k$, s induces the following posterior: the probability that the true state is in $\cup_{n=1}^{k} T_n$ is larger than $1 - c$, i.e.

$$\Pr(\theta \in \cup_{n=1}^{k} T_n | s) \geq 1 - c, \forall s \in S_k.$$

Finally, denote

$$T^* = \cup_{k=1}^{\infty} T_k.$$

Note that for every k, S_k, S^*, T_k, and T^* are π specific, and we use $S_k|\pi$, $S^*|\pi$, $T_k|\pi$, and $T^*|\pi$ to denote the corresponding sets under information policy when necessary.

Step 2. *We prove that a necessary and sufficient condition for the regime to survive is $\theta \in T^*$.*

We first show the sufficiency. If $\theta \in T^*$, there exists k such that $\theta \in T_k$ and $\theta \notin T_l$ for $l = 1, 2, ..., k-1$. We show that the regime survives for any $k = 1, 2, ...$ Suppose the agents coordinate on attacking if their signals are in S; then by the definition of T_1 and S_1, an individual agent whose signal is in S_1 would prefer to deviate to not attacking. By the rule of coordination, no agent shall attack if her signal is in S_1, and every $\theta \in T_1$ always survives under information policy $\pi(\cdot|\theta)$. By a similar argument, suppose the agents coordinate on attacking if their signals are in $S \backslash S_1$; then an individual agent whose signal is in S_2 would prefer to deviate to not attacking, and every $\theta \in T_1 \cup T_2$ always survives. The rest of the proof follows by mathematical induction.

We prove the necessity by contrapositive. First, by the proof of Proposition 1, every agent shall attack if and only if her signal realization is not in S^*. Then by the definition of T^*, for every state x not in T^*, the designer sends a signal in S^* with probability less than $1-x$; otherwise x is in T^*. Thus, every state x not in T^* is attacked by a mass greater than x and eventually fails. This completes the proof of the necessity.

Step 3. *We identify an upper bound of the ex ante probability that the regime survives, $\int_{T^*} f(\theta)d\theta$.*

Fix any information structure (S, π), and define a function $T : \Theta \to \mathbb{N}$ such that for every $\theta \in T^*$, we have $\theta \in \cup_{n=1}^{T(\theta)} T_n \backslash \cup_{n=1}^{T(\theta)-1} T_n$. By the definition of $T_{(\cdot)}$, $T(\theta)$ is unique for every θ. Intuitively, for every $\theta \in T^*$, $T(\theta)$ means that θ survives after and only after $T(\theta) - 1$ rounds of IESDS.

For $k = 1, 2, ...,$ define "discredit D_k":

$$D_k = \int_{\cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n} f(\theta) \int_{S_{k-1}} \pi(s|\theta) ds d\theta,$$

which is the measure of signals in S_{k-1} being sent when $\theta \in \cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n$. Similarly, for $k = 1, 2, \cdots, p = k+1, k+2, \cdots,$ define "credit $C_{k,p}$":

$$C_{k,p} = \int_{\cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n} f(\theta) \int_{S_p \backslash S_{p-1}} \pi(s|\theta) ds d\theta,$$

which is the measure of signals in $S_p \backslash S_{p-1}$ being sent when $\theta \in \cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n$. Intuitively, in each round of the IESDS, to deter a coordinated attack, a signal must induce a posterior belief that the true state is sufficiently likely to be strong; hence the chance of defeating it is sufficiently low.

Consider an arbitrary round k. The states that survive after the $k-1$th round of IESDS are $\cup_{n=1}^{k} T_n$; these states are considered as the strong states. A strong state discourages the agents from attacking the signals that it sends with positive probability, increasing the probability that the underlying true state is strong. Analogously, it's like providing "credit" to these signals. The amount of credit a strong state provides to a signal is the ex-ante probability measure that it sends this signal. Then $C_{k,p}$ is the aggregate credit all the states in $\cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n$ provide to the signals saved in the pth round of the IESDS, $S_p \backslash S_{p-1}$.

The weak states, however, are the states that still fail after the $k-1$th round of IESDS. A weak state encourages the agents to attack the signals that it sends with positive probability, decreasing the probability that the underlying true state is strong. Analogously, it's like drawing out credit from (or injecting "discredit" to) those signals. The amount of credit a weak state charges from a signal is the ex-ante probability measure that it sends this signal. Then D_{k+1} is the aggregate discredit all the states in $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$ charge from the signals saved in the previous rounds S_k.

In the kth round of the IESDS, to save the states in $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$, the information policy charges credit D_{k+1} from the signals saved in this and the previous rounds, S_k. The credit must not be overdrawn (specified later); otherwise those signals become too weak and the agents shall attack them in previous rounds. In the pth round of the IESDS, however, these states, $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$, are strong states; they provide credit $C_{k+1,p}$ to save the states in $\cup_{n=1}^{p+1} T_n \backslash \cup_{n=1}^{p} T_n$. In conclusion, in each round of the IESDS, the newly saved states "pollute" the strong signals endorsed by the states saved in the previous rounds; nevertheless it creates spaces for the states saved in the latter rounds to pollute.

The above intuition leads, for every round k, to two conditions that characterize an upper bound of the ex-ante probability that the regime survives: first, the credit must not be overdrawn; second, the states in $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$ are saved in the kth round. Precisely, consider round k. By the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for every $s \in S_k$, an individual agent receiving s shall not attack even if every agent receiving a signal not in S_{k-1} attacks; that is to say, if she attacks, the probability of winning is smaller than c. Then a necessary condition for T_n to be saved is

$$c \geq \frac{\sum_{m=1}^{k+1} D_m}{\sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p} + \sum_{m=1}^{k+1} D_m} \Leftrightarrow c \sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p} \geq (1-c) \sum_{m=1}^{k+1} D_m.$$

Also, by the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for $m = 1, 2, \cdots, k+1$, for every $\theta \in \cup_{n=1}^{m} T_n \backslash \cup_{n=1}^{m-1} T_n$, $\int_{S_{m-1}} \pi(s_i|\theta) ds_i \geq \min\{0, 1-\theta\}$; this further implies $\int_{S_k \backslash S_{m-1}} \pi(s_i|\theta) ds_i \leq \max\{1, \theta\}$.

Expanding the above condition yields

$$\begin{aligned}
& c \sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p} \geq (1-c) \sum_{m=1}^{k+1} D_m \\
\Leftrightarrow \;\; & c \sum_{m=1}^{k} \int_{\cup_{n=1}^{m} T_n \backslash \cup_{n=1}^{m-1} T_n} f(\theta) \int_{S_k \backslash S_{m-1}} \pi(s|\theta) ds d\theta \\
\geq \;\; & (1-c) \sum_{m=1}^{k} \int_{\cup_{n=1}^{m} T_n \backslash \cup_{n=1}^{m-1} T_n} f(\theta) \int_{S_m} \pi(s|\theta) ds d\theta \\
\Leftrightarrow \;\; & c \int_{\cup_{n=1}^{k} T_n} f(\theta) \int_{S_k \backslash S_{T(\theta)-1}} \pi(s|\theta) ds d\theta \\
\geq \;\; & (1-c) \int_{\cup_{n=1}^{k+1} T_n} f(\theta) \int_{S_{T(\theta)-1}} \pi(s|\theta) ds d\theta \\
\Rightarrow \;\; & c(\int_{T_1} f(\theta) d\theta + \int_{\cup_{n=2}^{k} T_n \backslash T_1} \theta f(\theta) d\theta) \geq (1-c) \int_{\cup_{n=1}^{k} T_n \backslash T_1} (1-\theta) f(\theta) d\theta
\end{aligned}$$

Next we introduce a lemma that helps us to focus on information policies that induce a specific form of agent equilibrium. Intuitively, it shows that the information designer can always construct an information policy π' that weakly improves upon π by, loosely speaking, saving the strong states.

Lemma 4. *For every information policy π, there exists another information policy π' such that $T^*|\pi' \supseteq (F^{-1}(1 - \int_{T^*|\pi} f(\theta) d\theta), \bar{\theta}]$.*

Now we are in the position to identify an upper bound of the ex ante probability that the regime survives, $\int_{T^*} f(\theta) d\theta$. Let $\tilde{\theta} = F^{-1}(1 - \int_{T^*} f(\theta) d\theta)$, by Lemma 4, as $k \to \infty$ we have

$$\begin{aligned}
& c(\int_{T_1} f(\theta) d\theta + \int_{\cup_{n=2}^{\infty} T_n \backslash T_1} \theta f(\theta) d\theta) \geq (1-c) \int_{\cup_{n=1}^{\infty} T_n \backslash T_1} (1-\theta) f(\theta) d\theta \\
\Rightarrow \;\; & c(\int_{1}^{\bar{\theta}} f(\theta) d\theta + \int_{\tilde{\theta}}^{1} \theta f(\theta) d\theta) - (1-c) \int_{\tilde{\theta}}^{1} (1-\theta) f(\theta) d\theta \geq 0
\end{aligned}$$

Suppose that π improves and $\int_{T^*} f(\theta) d\theta$ increases, $\tilde{\theta}$ decreases, the left-hand side of the second inequality above either always increases, or increases at first, then decreases. Thus, there exists a unique lower bound of $\tilde{\theta}$. If the lower bound is 0, there exists π such that every state survives; otherwise if the lower bound is strictly larger than 0, we use $\theta^{*\prime}$ to denote this lower bound, and $\theta^{*\prime}$ solves

$$c \int_{1}^{\bar{\theta}} f(\theta) d\theta + \int_{\theta^{*\prime}}^{1} (\theta + c - 1) f(\theta) d\theta \geq 0 \tag{10}$$

It's straightforward that $\theta^{*\prime}$ is unique, and then the upper bound of the ex ante probability that the regime survives is $1 - F(\theta^*)$.

Step 4. *We show that the probability of the status quo's survival under π^* exactly equals to the upper bound we proposed. π^* is therefore an optimal signal.*

As shown in the main text, the equilibrium outcome under π^* is that every agent who receives a signal in $\{s_k\}_{k=1}^{\infty}$ does not attack; as a result, the status quo survives whenever $\theta \in (\theta^*, \bar{\theta}]$. When receiving s_a, it is common knowledge that the state is in $[\underline{\theta}, \theta^*]$, so all agents attack, and the status quo is overthrown. Also, by the definition of T_k, under π^*, for $k = 1, 2, \cdots$, we have $T_k = (\theta_k, \theta_{k-1}]$.

Then by (1), (2), and (3)

$$\begin{aligned} c\left(\int_{\theta_1}^{\theta_0} f(\theta)d\theta + \sum_{k=3}^{\infty}\int_{\theta_{k-1}}^{\theta_{k-2}} \theta f(\theta)d\theta\right) &= (1-c)\sum_{k=2}^{\infty}\int_{\theta_k}^{\theta_{k-1}}(1-\theta)f(\theta)d\theta \\ c\left(\int_{\theta_1}^{\theta_0} f(\theta)d\theta + \int_{\theta^{*\prime}}^{\theta_1} \theta f(\theta)d\theta\right) &= (1-c)\int_{\theta^{*\prime}}^{\theta_1}(1-\theta)f(\theta)d\theta \end{aligned}$$

Notably, θ^* indeed solves (10); as the solution is unique, we have $\theta^* = \theta^{*\prime}$. The measure of $\int_{T^*} f(\theta)d\theta$ exactly equals the upper bound we proposed; thus θ^* is optimal.

Lastly, all the steps above assume that not every state survives under π^*. If otherwise, for some k we have $\theta_k < \underline{\theta}$, then π^* has already achieved the best outcome possible and is thus optimal.

A.3 Miscellaneous Proofs

Proof of Lemma 4. Fix any π, construct π' as follows: as $\tilde{\theta}$ decreases from $\bar{\theta}$, for every $\tilde{\theta}$, let

$$\int_{T^* \cap (\tilde{\theta}, \bar{\theta}]} f(\theta)\pi(\cdot|\theta)d\theta \equiv \int_{F^{-1}(1-\int_{T^* \cap (\tilde{\theta},\bar{\theta}]} f(u)du)}^{\bar{\theta}} f(\theta)\pi'(\cdot|\theta)d\theta$$

and

$$\int_{(\tilde{\theta},\bar{\theta}] \setminus T^*} f(\theta)\pi(\cdot|\theta)d\theta \equiv \int_{F^{-1}(1-\int_{(\tilde{\theta},\bar{\theta}] \cup T^*} f(v)dv)}^{F^{-1}(1-\int_{T^*} f(w)dw)} f(\theta)\pi'(\cdot|\theta)d\theta$$

As $f(\theta) \leq 1$ for every θ, such a π' always exists.

Intuitively, fix any π, the states that survive are T^* and the ex-ante probability that the regime survives is $\int_{T^*|\pi} f(\theta)d\theta$. Under information policy π', the top $\int_{T^*|\pi} f(\theta)d\theta$ states and the states in T^* switch their signal distributions.

Next we prove that every state in $(F^{-1}(1 - \int_{T^*|\pi} f(\theta)d\theta), \bar{\theta}]$ survives under information policy π'. Consider any agent i, upon receiving any signal s_i, under π', her posterior belief on the state set $(F^{-1}(1 - \int_{T^*} f(w)dw, \bar{\theta}] \cup T^*$ is the same as under π. Then consider any second-order infinitely small probability measure that any state $\theta \in T^*$ takes up in the prior under π, there exists a corresponding second-order infinitely small probability measure that some state $\theta' \in (F^{-1}(1 - \int_{T^*} f(w)dw, \bar{\theta}]$ takes up in the prior; θ' induces exactly the same signal distribution under π' as what θ induces under π, and $\theta' \geq \theta$. Fix θ, any potential coordination that is not self-sustainable (i.e., some agents shall unilaterally deviate) under π is weakly less likely to successfully overthrow a regime of state θ' under π'; therefore, such a coordination yields weakly less expected payoff for every agent and is not self-sustainable under π'. Similarly, any coordination plan that fails to overthrow a regime of state θ under π also fails to overthrow a regime of state θ' under π'. As θ survives under π, θ' survives under π'. Then as every state in T^* survives under π, every state in $(F^{-1}(1 - \int_{T^*|\pi} f(\theta)d\theta), \bar{\theta}]$ survives under π'. □

Proof of Proposition 2.B. The first statement is straightforward.

To prove the second statement, rewrite (3) for F and G to get

$$c(1 - F(\theta^*)) = \int_{\theta^*}^{1} (F(\theta) - F(\theta^*))d\theta$$
$$c(1 - G(\theta^{**})) = \int_{\theta^{**}}^{1} (G(\theta) - G(\theta^{**}))d\theta.$$

Consider θ' such that $G(\theta') = F(\theta^*)$ which implies that $\theta' \geq \theta^*$ by first-order stochastic dominance. As $G(\theta) \leq F(\theta)$ for all θ, we know that $\int_{\theta'}^{1}(G(\theta) - G(\theta'))d\theta \leq \int_{\theta^*}^{1}(F(\theta) - F(\theta^*))d\theta$, i.e. $c(1 - G(\theta')) \geq \int_{\theta'}^{1}(G(\theta) - G(\theta'))d\theta$. As the left-hand side of (3) must be negative for all $\theta < \theta^{**}$ and positive for all $\theta > \theta^{**}$, it must be that $\theta^{**} \leq \theta'$. Therefore $1 - G(\theta^{**}) \geq 1 - G(\theta') = 1 - F(\theta^*)$.

To prove the third statement, observe from (3) that a sufficient condition for $1 - F(\theta^*)$ to be bounded away from 0 is that the measure of $\theta \in [1 - c, 1]$ is bounded away from 0. This is ensured by $f(\theta) > 0$ for all $\theta \in \hat{\Theta}$, for some non-empty $\hat{\Theta} \subset [1 - c, 1]$. The result thus follows. □

Proof of Proposition 3. We first consider increasing c. Note that $\theta^\dagger$ and θ^* are characterized by

$$c(F(\bar{\theta}) - F(\theta^\dagger)) - (F(1) - F(\theta^\dagger)) = 0 \tag{11}$$
$$c(F(\bar{\theta}) - F(\theta^*)) - (F(1) - F(\theta^*)) + \int_{\theta^*}^{1} \theta f(\theta)d\theta = 0, \tag{12}$$

where (12) is a representation of (3). (11)-(12) gives

$$(1 - c)(F(\theta^\dagger) - F(\theta^*)) = \int_{\theta^*}^{1} \theta f(\theta)d\theta$$
$$(1 - c)(F(\theta^\dagger) - F(\theta^*)) = \int_{\theta^*}^{1} \theta f(\theta)d\theta$$
$$F(\theta^\dagger) - F(\theta^*) = \frac{\int_{\theta^*}^{1} \theta f(\theta)d\theta}{1 - c}$$

It is clear that θ^* decreases as c increases. Hence $F(\theta^\dagger) - F(\theta^*)$ increases as c increases.

As $F(1) \to 1$, $F(\theta^\dagger) \to 1$. Then if $F(\theta^*) \to 1$ we have $\int_{\theta^*}^1 \theta f(\theta) d\theta \to 1$. Then we require $0 = \frac{1}{1-c}$, contradiction. Thus θ^* is bounded away from 1. □

B Appendix: Proofs of Robustness (For Online Publication)

Proof of Proposition 4. Suppose that π'_n is an optimal policy and π'_n is different from π_n. Through the following analysis we assume $\theta_k \geq \underline{\theta}$ for every $k \leq n$.

It's straightforward that n signal realizations can induce at most $n-1$ rounds of IESDS. Suppose that π'_n induces $m \leq n-1$ rounds of IESDS.

For π_n, we have

$$\begin{aligned} D_1|\pi_n &= 0, C_{1,1}|\pi_n = \bar{\theta} - 1 \\ D_2|\pi_n &= \frac{c}{1-c} C_{1,1}|\pi_n \\ D_2|\pi_n + D_3|\pi_n &= \frac{c}{1-c}(C_{2,2}|\pi_n + C_{1,1}|\pi_n) \\ &\cdots \\ \sum_{p=2}^{n} D_p|\pi_n &= \frac{c}{1-c}\sum_{p=2}^{n} C_{p-1,p-1}|\pi_n. \end{aligned}$$

For π'_n, we have

$$\begin{aligned} D_1|\pi'_n &= 0 \\ D_2\pi'_n &\leq \frac{c}{1-c} C_{1,1}|\pi'_n \\ D_2|\pi'_n + D_3|\pi'_n &\leq \frac{c}{1-c}(C_{2,2}|\pi'_n + C_{1,1}|\pi'_n + C_{1,2}|\pi'_n) \\ \sum_{p=2}^{4} D_p|\pi'_n &\leq \frac{c}{1-c}(C_{3,3}|\pi'_n + \sum_{p=2}^{3} C_{2,p}|\pi'_n + \sum_{p=1}^{3} C_{1,p}|\pi'_n) \\ &\cdots \\ \sum_{p=2}^{m+1} D_p|\pi'_n &\leq \frac{c}{1-c}(C_{m,m}|\pi'_n + \sum_{p=m-1}^{m} C_{m-1,p}|\pi'_n + ... + \sum_{p=1}^{m} C_{1,p}|\pi'_n). \end{aligned}$$

If $C_{1,1}|\pi'_n < C_{1,1}|\pi_n$, then $D_2|\pi'_n < D_2|\pi_n$, then $C_{2,2}|\pi'_n < C_{2,2}|\pi_n$, also we know $C_{1,2}|\pi'_n + C_{1,1}|\pi'_n \leq C_{1,1}|\pi_n$, then $D_2|\pi'_n + D_3|\pi'_n < D_2|\pi_n + D_3|\pi_n$, then $C_{3,3}|\pi'_n < C_{3,3}|\pi_n, \cdots$ following a mathematical induction we have $\sum_{p=2}^{m+1} D_p|\pi'_n < \sum_{p=2}^{m+1} D_p|\pi_n$, as $m \leq n-1$, $\sum_{p=2}^{m+1} D_p|\pi'_n \leq \sum_{p=2}^{n} D_p|\pi_n$.

By the proof of Theorem 1, step 3, under any information policy, the minimum discredit a state θ that survives charges is $1-\theta$. Thus, fix $\sum_{p=2}^{n} D_p|\pi_n$, $\cup_{p=1}^{n} T_n|\pi_n$ uniquely maximizes the information designer's ex ante probability of survival. As

$\sum_{p=2}^{m+1} D_p|\pi'_n < \sum_{p=2}^{n} D_p|\pi_n$, the information designer's ex ante probability of survival under π'_n is strictly smaller than under π_n, and we reach a contradiction. Thus, in every optimal design, $C_{1,1}|\pi'_n = C_{1,1}|\pi_n$, $D_2|\pi'_n = D_2|\pi_n$; then we also have $C_{1,p}|\pi'_n = 0$ for $p = 2, 3, \cdots, m$.

Similarly, given $C_{1,1}|\pi'_n = C_{1,1}|\pi_n$, $D_2|\pi'_n = D_2|\pi_n$, suppose that $C_{2,2}|\pi'_n < C_{2,2}|\pi_n$, then $D_3|\pi'_n < D_3|\pi_n$, then $C_{3,3}|\pi'_n < C_{3,3}|\pi_n$, also we know $C_{2,3}|\pi'_n + C_{2,2}|\pi'_n \leq C_{2,2}|\pi_n$, then $D_3|\pi'_n + D_4|\pi'_n < D_3|\pi_n + D_4|\pi_n$, then $C_{4,4}|\pi'_n < C_{4,4}|\pi_n, \cdots$ following a mathematical induction we have $\sum_{p=3}^{m+1} D_p|\pi'_n < \sum_{p=3}^{m+1} D_p|\pi_n \leq \sum_{p=3}^{n} D_p|\pi_n$. Note that we already have $D_2|\pi'_n = D_2|\pi_n$, thus we have $\sum_{p=2}^{m+1} D_p|\pi'_n < \sum_{p=2}^{n} D_p|\pi_n$, by the same argument as above, the information designer's ex ante probability of survival under π'_n is strictly smaller than under π_n, and we reach a contradiction. Thus, in every optimal design, $C_{2,2}|\pi'_n = C_{2,2}|\pi_n$, $D_3|\pi'_n = D_3|\pi_n$, then we also have $C_{2,p}|\pi'_n = 0$ for $p = 3, 4, \cdots, m$.

Iterate the above process, by a mathematical induction, we conclude that in every optimal design, for $p = 1, 2, \cdots, m$, $C_{p,p}|\pi'_n = C_{p,p}|\pi_n$, $D_{p+1}|\pi'_n = D_{p+1}|\pi_n$, and $C_{p,q}|\pi'_n = C_{p,q}|\pi_n = 0$ for $q = p+1, p+2, \cdots, m$.

By the above analysis, in every optimal design π'_n, states in $T_1|\pi'_n$ send signals in $S_1|\pi'_n$ with probability 100%, and states in $\Theta \backslash (T_1|\pi'_n \cup T_2|\pi'_n)$ should not send any signal in $S_1|\pi'_n$ with positive probability. We show that if S_1 contains more than one element (denoted by s_1), it must not be optimum. First, we construct information policy π''_n; under this information policy, the states in $\Theta \backslash (T_1|\pi'_n \cup T_2|\pi'_n)$ behave the same as under π'_n, the states in $T_1|\pi'_n \cup T_2|\pi'_n$ send s_1 whenever they should send a signal in S_1 under π'_n. It's straight forward that π''_n uses at least one signal less than π'_n does, and the outcome is the same as π'_n. Then we can construct information policy π'''_n that improves upon π''_n by using one more signal, denoted by s'. Under π'''_n, let the states in T_n send s' whenever they should send a signal not in S_{n-1} under π''_n, then let a sufficiently small state set in $\Theta \backslash (\cup_{p=1}^{m} T_p|\pi''_n)$ send s' with probability 100%, the other states behave the same as under π''_n. All the states that survive under π''_n still survive under π'''_n; and the small state set that sends s' with probability 100% now survives. Thus the ex ante probability of survival under π'''_n is strictly larger than under π''_n; this contradicts our assumption that π'_n is optimum. Thus, states in $T_1|\pi'_n$ send s_1 with probability 100%, and a state in $T_2|\pi'_n$ sends s_1 with probability that equals to its strength.

By similar arguments, in every optimal design, a state in $T_2|\pi'_n$ sends one single signal, s_2, with probability that equals to one minus its strength. This iteration proceeds, and we show that in every optimal design, $\pi'_n = \pi_n$. □

Proof of Corollary 1. Fix any information policy $\pi(\cdot|\theta)$; we define a series $T_{(\cdot)}$ on the state space and, for every i, a series $S^i_{(\cdot)}$ on the signal space. The idea is similar to the proof of Proposition 1 and the proof of Theorem 1, with minor modifications as follows.

The difference between this specification and Theorem 1 is that now, for each state, the ex post distribution of signals is not determinate; instead, it can be any distribution over all possible ex post distributions. Nevertheless, we show that this added degree of freedom does not increase the maximum of the status quo's probability of survival, i.e. the original design remains optimal.

For every i, define $S_0^i = \emptyset$ and $a_i^0(s) \equiv 1$. Define $S_1^i \subseteq S$ as the set of states satisfying the following condition:

$$\int_{\Theta} \frac{f(\theta')\pi(s|\theta')}{\int_{\Theta} f(\theta'')\pi(s|\theta'')d\theta''} \Pr(\theta' < \int_{[0,1]} a_j^0(s_j)dj|\theta = \theta', s_i = s)d\theta' \leq c.$$

Define

$$a_i^1(s) = \begin{cases} 0 \text{ if } s \in S_1^i \\ 1 \text{ otherwise.} \end{cases}$$

We then perform an iteration similar to the proof of Proposition 1. For every i and for $k = 2, 3, ..., +\infty$, we define a_i^k, S_k^i, S^{i*}. We omit the proof that $\{S^{i*}\}_{i\in[0,1]}$ characterizes the unique agent equilibrium.

To save notations and to simplify our discussion, we apply the following transformation on the signal space and, correspondingly, on π. Notice that for any i and any point-to-point transformation A from S to S, let π_i' be an alternative signal such that for every $\theta \in \Theta, s \in S$, $\pi_i'(A(s)|\theta) = \pi_i(s|\theta)$, the payoff yielded by $\pi' = \{\pi_i'\}_{i\in[0,1]}$ is the same as π. Thus, we can always construct a proper transformation A to attain arbitrary S_k^i for every i, k, and the status quo's ex ante probability of survival is unchanged. Then, for every $\pi(\cdot|\theta)$, we can always find $\pi'(\cdot|\theta)$ such that $\bar{a}_i^k(\cdot) \equiv \bar{a}_j^k(\cdot)$ for every k and every $i, j \in [0,1]$, and the status quo's ex ante probability of survival under $\pi'(\cdot|\theta)$ is identically equal to its ex ante probability of survival under $\pi(\cdot|\theta)$.

Without loss of generality, from here on, we focus on policies under which $\bar{a}_i^k(\cdot) \equiv \bar{a}_j^k(\cdot)$ for every n and every $i, j \in [0,1]$, then $\bar{S}_k^i = \bar{S}_k^j$ for every k and every $i, j \in [0,1]$. To save notations, we still call this policy $\pi(\cdot|\theta)$, and define $\bar{a}^k(\cdot) \equiv \bar{a}_i^k(\cdot)$, $S_k = S_k^i$, for arbitrary i and for every k.

Next, define type set $T_0 = \emptyset$ and function $f_0(\theta)$ on Θ, $f_0(\theta) = 0$ for every $\theta \in \Theta$.

Define type set $T_1 = (1, \bar{\theta}]$ and function $f_1(\theta)$ on Θ, $f_1(\theta) = f(\theta)$ for every $\theta \in T_1$ and $f_1(\theta) = 0$ elsewhere.

Define type set T_2 as every state $x \in \Theta$ such that: $\Pr(\int_{i\in[0,1]} \Pr(s_i \in S_1^i|\theta = x)di < x) > 0$; define function $f_2(\theta)$ on Θ, for every θ, $f_2'(\theta) = f(\theta)\Pr(\int_{i\in[0,1]} \Pr(s_i \in S_1^i|\theta = x)di < x)$.

Define T_k and $f_k(\theta)$ for $k = 3, 4, ...$ similarly.

Next, we identify an upper bound of the ex ante probability that the regime survives.

Fix any information policy $\pi(\cdot|\theta)$ that induces a unique agent equilibrium which satisfies the above condition, define function $T(\theta)$ the same as in the proof of Theorem 1.

For $i \in [0,1], k = 1, 2, ...,$ define "discredit D_k":

$$D_k^i = \int_\Theta [f_k(\theta) - f_{k-1}(\theta)] \Pr(s_i \in S_{k-1}|\theta) d\theta.$$

For $i \in [0,1], k = 1, 2, ..., p = k+1, k+2, ...,$ define "credit $C_{k,p}^i$":

$$C_{k,p}^i = \int_\Theta [f_k(\theta) - f_{k-1}(\theta)] \Pr(s_i \in S_p \backslash S_{p-1}|\theta) d\theta.$$

By the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for every k, i, $c \sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p}^i \geq (1-c) \sum_{m=1}^{k+1} D_m^i$. Also note that for the status quo to survive it must send signals in S_k to a population greater or equal to $1-\theta$. Thus, similar to the proof of Theorem 1, for every i, θ

$$\begin{aligned} c \int_\Theta f_k(\theta) \Pr(s_i &\in S_k \backslash S_{T(\theta)-1}|\theta) d\theta \\ &\geq (1-c) \int_\Theta f_{k+1}(\theta) \Pr(s_i \in S_{T(\theta)-1}) d\theta \end{aligned}$$

Thus we have

$$\begin{aligned} & c \int_{[0,1]} \int_\Theta f_k(\theta) \Pr(s_i \in S_k \backslash S_{T(\theta)-1}|\theta) d\theta di \\ \geq & (1-c) \int_{[0,1]} \int_\Theta f_{k+1}(\theta) \Pr(s_i \in S_{T(\theta)-1}|\theta) d\theta di \\ \Rightarrow & c(\int_\Theta f_1(\theta) d\theta + \int_\Theta (f_k(\theta) - f_1(\theta)) \theta f(\theta) d\theta) \geq \int_\Theta f_k(\theta)(1-\theta) d\theta \end{aligned}$$

Note that Lemma 4 remains valid; the policy maker can still construct an information policy that weakly improves upon π by saving the high types. Thus in any optimum, $f_k(\theta) = f(\theta)$ for $\theta \in T_k$ and $f_k(\theta) = 0$ elsewhere. Following the proof of Theorem 1, θ^* satisfies

$$c(\int_1^{\bar{\theta}} f(\theta) d\theta + \int_{\theta^*}^1 \theta f(\theta) d\theta) = (1-c) \int_{\theta^*}^1 (1-\theta) f(\theta) d\theta$$

which is identical to the proof of Theorem 1. Notice that the last condition is irrelevant to agent identity i. Thus from here on, the rest of the proof follows the proof of Theorem 1. □

References

Acemoglu, D. and J. A. Robinson (2005). *Economic origins of dictatorship and democracy*. Cambridge University Press.

Alder, S. and Y. Wang (2019). Divide and rule: How political elites polarize society. *University of Essex and University of North Carolina*.

Alonso, R. and O. Câmara (2016). Persuading voters. *American Economic Review 106*(11), 3590–3605.

Angeletos, G.-M., C. Hellwig, and A. Pavan (2006). Signaling in a global game: Coordination and policy traps. *Journal of Political Economy 114*(3), 452–484.

Bail, C. A., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221.

Banerjee, A. V. (1993). The economics of rumours. *Review of Economic Studies 60*(2), 309–327.

Bardhi, A. and Y. Guo (2018). Modes of persuasion toward unanimous consent. *Theoretical Economics 13*(3), 1111–1149.

Basak, D. and Z. Zhou (2018). Timely persuasion. working paper.

Basak, D. and Z. Zhou (2019). Diffusing coordination risk. *American Economic Review* (forthcoming).

Bergemann, D. and S. Morris (2016). Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics 11*(2), 487–522.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Best, J. and D. Quigley (2017). Persuasion for the long run. *Available at SSRN 2908115*.

Carlsson, H. and E. van Damme (1993). Global games and equilibrium selection. *Econometrica 61*(5), 989–1018.

Chan, J., S. Gupta, F. Li, and Y. Wang (2019). Pivotal persuasion. *Journal of Economic Theory 180*, 178–202.

Chassang, S. and G. P. I. Miquel (2010). Conflict and deterrence under strategic risk. *The Quarterly Journal of Economics 125*(4), 1821–1858.

Chen, H., Y. K. Lu, and W. Suen (2016). The power of whispers: A theory of rumor, communication, and revolution. *International Economic Review 57*(1), 89–116.

Cong, L. W., S. R. Grenadier, and Y. Hu (2019). Dynamic interventions and informational linkages. *Journal of Financial Economics* (forthcoming).

Dai, L. and M. Yang (2018). Organizations and coordination in a diverse population. *Available at SSRN 3051192*.

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

Edmond, C. and Y. K. Lu (2017). Creating confusion: Information manipulation and imperfect coordination. working paper.

Ely, J. C. (2017). Beeps. *American Economic Review 107*(1), 31–53.

Galperti, S. and J. Perego (2018). A dual perspective on information design. *Available at SSRN 3297406*.

Galperti, S. and J. Perego (2019). Belief meddling in social networks: An information-design approach. *Available at SSRN 3340090*.

Glaeser, E. L. (2005). The political economy of hatred. *The Quarterly Journal of Economics 120*(1), 45–86.

Goldstein, I. and C. Huang (2016). Bayesian persuasion in coordination games. *American Economic Review: Papers & Proceedings 106*(5), 592–596.

Goldstein, I. and C. Huang (2018). Credit rating inflation and firms' investments. *Available at SSRN 3082428*.

Guo, Y. and E. Shmaya (2019). Costly miscalibration. Technical report, Northwestern, working paper.

Guriev, S. and D. Treisman (2019). Informational autocrats. *Journal of Economic Perspectives 33*(4), 100–127.

Hébert, B. and M. Woodford (2017). Rational inattention and sequential information sampling. Technical report, National Bureau of Economic Research.

Hoshino, T. (2019). Multi-anent persuasion: Leveraging strategic uncertainty. *ITAM*.

Huang, C. (2017). Defending against speculative attacks: The policy maker's reputation. *Journal of Economic Theory 171*, 1–34.

Inostroza, N. and A. Pavan (2018). Persuasion in global games with application to stress testing. working paper, Northwestern University and University of Toronto.

Jeitschko, T. D. and C. R. Taylor (2001). Local discouragement and global collapse: a theory of coordination avalanches. *American Economic Review 91*(1), 208–224.

Jeong, D. (2019). Using cheap talk to polarize or unify a group of decision makers. *Journal of Economic Theory 180*, 50–80.

Kajii, A. and S. Morris (1997). The robustness of equilibria to incomplete information. *Econometrica: Journal of the Econometric Society*, 1283–1309.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

King, G., J. Pan, and M. E. Roberts (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review 111*(3), 484–501.

Levy, G. and R. Razin (2018). Information diffusion in networks with the bayesian peer influence heuristic. *Games and Economic Behavior 109*, 262–270.

Li, W. and X. Tan (2019). Locally bayesian learning in networks. *Theoretical Economics* (forthcoming).

Lipnowski, E., D. Ravid, and D. Shishkin (2018). Persuasion via weak institutions. *Available at SSRN 3168103*.

Lipnowski, E. and E. Sadler (2019). Peer-confirming equilibrium. *Econometrica 87*(2), 567–591.

Mathevet, L., D. Pearce, and E. Stacchetti (2018). Reputation and information design. *New York University*.

Mathevet, L., J. Perego, and I. Taneva (2019). On information design in games. *Journal of Political Economy, forthcoming*.

Morris, S., D. Oyama, and S. Takahashi (2019). Adversarial information design in binary-action supermodular games. *MIT*.

Morris, S., R. Rob, and H. S. Shin (1995). p-dominance and belief potential. *Econometrica: Journal of the Econometric Society*, 145–157.

Morris, S. and H. S. Shin (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review 88*(3), 587–597.

Morris, S. and H. S. Shin (2003). Global games: Theory and applications. In M. Dewatripont, L. P. Hansen, and S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Volume 1, Cambridge. Cambridge University Press.

Morris, S. and M. Yang (2019). Coordination and continuous stochastic choice. Technical report, Duke University.

Olson, M. (1965). *The logic of collective action*, Volume 124. Harvard University Press.

Ong, J. C. and J. Cabanes (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the philippines. *Newton Tech4Dev Network*.

Pomatto, L., P. Strack, and O. Tamuz (2018). The cost of information. *arXiv preprint arXiv:1812.04211*.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under" almost common knowledge". *The American Economic Review*, 385–391.

Shadmehr, M. and D. Bernhardt (2011). Collective action with uncertain payoffs: coordination, public signals, and punishment dilemmas. *American Political Science Review 105*(4), 829–851.

Shadmehr, M. and D. Bernhardt (2015). State censorship. *American Economic Journal: Microeconomics 7*(2), 280–307.

Taneva, I. (2019). Information design. *American Economic Journal: Microeconomics, forthcoming*.