The Power of Whispers: A Theory of Rumor, Communication and Revolution*

HENG CHEN
University of Hong Kong
YANG K. LU
Hong Kong University of Science and Technology
WING SUEN
University of Hong Kong

Abstract. We study the role rumors play in revolutions using a global game model. Agents with diverse private information rationally evaluate the informativeness of rumors about the regime strength. Without communication among agents, wild rumors are discounted and agents are generally less responsive to rumors than to trustworthy news. When agents can exchange views on their assessment of rumors, a rumor against the regime would coordinate a larger mass of attackers if the regime is indeed weak. The effect of communication can be so large that rumors can have a greater impact on mobilization than does fully trustworthy information.

Keywords. coordination game, public signals, swing population, mixture distribution, censorship

JEL Classification. D74, D83

*Preliminary and Incomplete. PLEASE DO NOT POST IT ON THE INTERNET OR CIRCULATE.

1. Introduction

Collective actions, such as riots, protests and political campaigns, are often immersed in rumors. Perhaps the most dramatic theater to witness rumors in action is a political revolution. Amid the recent Tunisian revolution, Ben Ali, the ex-Tunisian leader, was said to have fled his country. This was confirmed after conflicting rumors about his whereabouts, and finally led to the end of street protests. A while later in Egypt, it was widely reported that Mubarak's family had left for London, which was believed by many as a clear sign of fragility of the regime. Similar rumors about Qaddafi and his family appeared in Libya when the battle between the opposition and the regime intensified. Rumors are not unique to the series of revolutions in the Arab Spring. During the 1989 democracy movement in China, rumors repeatedly surfaced about the death of the leaders, Deng Xiaoping and Li Peng, as well as the divide among communist leaders, all of which indicated the vulnerability of the regime.[1]

Are rumors just rumors? One might conjecture that rumors that spread during turmoils would quickly disappear without leaving a trace, as individuals might simply ignore, disbelieve, or at least discount unreliable information. However many historical incidents suggest otherwise. Rumors often turned out to be effective in mobilizing citizens to take action. The news about Mubarak's family proved to be false a while later, yet the opposition credited it for "mark[ing] a new phase" in their campaign (World-Tribune.com 2011). Another famous example is the Velvet Revolution in Czechoslovakia, which was described as a "revolution with roots in a rumor" (Bilefsky 2009). At the dawn of the revolution in 1989, a prominent rumor that a 19-year old college student was brutally killed by the police triggered many otherwise hesitant citizens to take to the streets. The revolution gained momentum right after that. Chinese history also offers anecdotes in which rumors mobilized mass participation in the Boxer Uprising, the Republican Revolution, and the May Fourth Movement.[2] Similarly, riots are often sparked by rumors as well: the 1921 Tulsa race riot and the 1967 Newark riots provide dramatic examples.

A common interpretation of the role of rumors in mass movements is that individuals are just blindly herded by rumors. In this paper, however, we adopt the position that individuals are fully aware that rumors circulating in times of turmoil may or may not be well founded, and that they update their beliefs in a Bayesian manner. Since rumors are widely circulated and commonly observed, they may serve as a coordi-

[1] There were widespread rumors of many variants that Deng died of illness during the protest and that Li was gun shot to death. It was also widely rumored that some senior leaders in the Communist Party wrote an open letter to oppose any actions against students. Some of these rumors were repeated in the print media. See, for example, the news reports in the daily newspaper *Ming Pao* on June 6, 1989.

[2] Zhang (2009) documents these events in details, supplemented by several contemporary cases in China.

nation device just like a public signal in a coordination game. We explain why some rumors are effective in mobilizing participation in collective actions while others are simply ignored, which depends on the extent to which rumors are deemed plausible by individuals with diverse private information.

Furthermore we emphasize how communication can reinforce the role of rumors. Individuals in times of uncertainty and crisis often seek others' opinions and discuss with peers about their judgment and evaluation of rumors. Information from fellow participants can influence their beliefs and even actions, especially when they are unsure of what actions they should take. We show that such "idle talk" about rumors between individuals could potentially overcome their skepticism and strengthen the impact of rumors on mobilization. With communication, it is possible that rumors can have a greater impact on mobilization than does fully trustworthy information.

Specifically, we model political revolution as a coordination game among a large number of citizens, who decide whether to revolt against the regime or not. Citizens are uncertain about the regime's strength and possess dispersed information on it. In this model, global strategic complementarities arise, i.e., citizen's incentive to revolt increases with aggregate action of all the other citizens. If the number of participants is sufficiently high, the regime collapses; otherwise it survives. Before citizens take actions, they hear a rumor about the strength of the regime. This rumor is a publicly observed message, which could be either an informative signal on the regime's strength or an uninformative noise unrelated to fundamentals. Citizens assess the informativeness of the rumor based on their own private information on the regime. As a consequence of diverse private information among citizens, their assessments may also differ. Citizens are also allowed to communicate with one another and exchange their assessments on the rumor. In other words, they tell each other whether they think the rumor is believable or not.

In this model, citizens understand that rumors could be uninformative and therefore remain skeptical of them. They make an inference on how likely a rumor is informative based on their own private information, using standard Bayesian rules. The likelihood they assign that the rumor is informative is endogenous: if the rumor is far different from what the citizen personally knows about the regime, she tends to discount it more heavily. One obvious implication of this mechanism is that very extreme rumors—news that almost no citizens would believe—have little effects on equilibrium outcomes. Not surprisingly, due to their skepticism, citizens are less responsive to the news they hear: rumors against the regime mobilize less attackers, compared to the hypothetical case where such news is known to be trustworthy.

When citizens are allowed to communicate with each other, a fraction of the popu-

lation (those with intermediate private information) will decide to attack the regime or not, depending on whether or not their peers believe in the public rumor. With communication citizens on the whole are better informed about whether the rumor they hear is close to the underlying state than they are without communication. Specifically, when the rumor is close to the true regime strength, more citizens will believe it is sufficiently informative and send confirmatory messages to their peers. Citizens who receive confirmatory messages believe that the regime strength is more likely to be close to what the rumor indicates, and therefore assign higher weight to states of nature close to the rumor. As a consequence, there will be more attackers than when communication is not allowed, if the rumor is against the regime and happens to be near the true state. Moreover the coordination effect of communication can be so strong that more regime attackers can be mobilized by a negative rumor in this communication model than in a model in which all citizens trust the news fully without any information exchange.

Although we assume that rumors are exogenous in this paper, our framework allows us to analyze the effects of alternative communication devices and the management of information by the state. First, the mode of communication matters for equilibrium outcomes. Specifically we show that if citizens can fully communicate their private information instead of just exchanging messages about whether or not they believe in the rumor, then they tend to rely less on the public information conveyed by the rumor. Second, sentiment—the public's perception of what untrustworthy news would sound like—is a double-edged sword. On the one hand, the regime could manipulate perceptions and discredit negative information against itself, so as to increase the likelihood of its survival. On the other hand, systematic propaganda attempts by the government to spread news about its alleged strength only makes rumors about the regime's weakness more credible. Third, we also use our framework to study the effects of censorship. Suppose the government blocks any public information against the regime and only allow positive rumors to circulate. Then citizens would interpret the absence of rumors as a sign of bad news for the regime. We can show that censorship in this model can backfire in that it does not necessarily help the regime to survive.

While not dismissing the relevance of the role of rumormongers and the process of how rumors spread during revolutions, our paper focuses on the effects of rumors and communication on the equilibrium in a coordination game, taking the origin and the content of such rumors as exogenous. Studies on rumors show that they could be created both intentionally or unintentionally (Knapp 1944, Nkpa 1977, Ley 1997, Grunden, Walker, and Yamazaki 2005, Elias and Scotson 1994, and Gambetta 1994). The incentives of rumormongers could be unpredictable in the sense that they might

be created neither for the purpose of mobilizing collective action nor for defending the regime (Zhang 2009 and Turner, Pratkanis, Probasco, and Leve 1992). Misunderstanding between individuals is a usual source of rumors (Allport and Postman 1947, Peterson and Gist 1951, and Buchner 1965). See, however, Edmond (2011) for a model of state information manipulation in a global game context. We also choose to abstract from how rumors travel from one to another.[3] It is implicitly assumed that rumors could reach to every citizen in the game.[4] This assumption seems to be not unrealistic for many revolutions in history: Rumors against authorities did gain a substantial, even huge, amount of publicity among individuals under very repressive regimes.[5]

This paper should not be interpreted as contradicting the literature that stresses structural factors as root causes for a revolution (Skocpol 1979). Structural factors, such as the state of the economy and international pressure, are those that make a society "ripe" for revolution. However, it has been noted that structural factors are not sufficient for a successful revolution. In line with Bueno de Mesquita (2010), we argue that some random factors also play a role in determining the fate of a revolution. In our model, the realization of rumors serves as a source of randomness.[6]

Our work enriches the global game literature in a couple of directions. Unlike most global game models, we offer a more general specification of public signals and allow citizens to be skeptical of the public information they observe. A common implicit assumption in the literature is that citizens believe the public signal is informative. When they form the posterior belief on the fundamental, citizens assign a constant weight on the public signal based on the its relative precision to private signals. The implication is that citizens would not adjust the weight, even though the public signal is remarkably different from what their own private information implies. However, such updating rule seems to be counter-intuitive, in view of the fact that individuals tend to disbelieve information which is too different from their priors. Our model provides a formal justification using mixture distributions to explain why the standard linear updating rule may not be appropriate.

[3]Sociological studies find two types of rumor propagation mechanisms: "snow balling pattern" (Peterson and Gist (1951)) and "simplification pattern" (Allport and Postman 1947 and Buchner 1965). Modeling the process of diffusion of rumors in a non-reduced form deserves a separate endeavor.

[4]In principal, we could also assume that a certain fraction of citizens do not hear any rumor. This would not affect the main results that we characterize in our model.

[5]The rumor that a college student was killed by the police, which ignited the Velvet Revolution in Czechoslovakia, was broadcast by Radio Free Europe. Modern mass media including the internet offer additional channels for spreading rumors. For example, in 2009, the Iranian post-election anti-government protest intensified after a rumor surfaced in the internet that police helicopters were pouring acid and boiling water on protesters (Esfandiari 2010).

[6]In contrast to the literature with multiple-equilibria, our model features unique equilibrium. We attribute the fundamental cause of success or failure of a revolution to the regime's strength. However, for the same strength of a regime, it could collapse when the realization of the rumor takes certain values, while it would survive when it takes some other values.

It seems that the coordination role of direct communication between individuals has been largely overlooked in games with large population. Typically, citizens are assumed to only respond to signals they observe, and further interaction between citizens are often left out for the sake of simplicity and tractability.[7] In reality, individuals do exchange information with each other before they make decisions and take actions. This is especially true in collective actions such as protests, demonstrations and revolutions. In our work, we model direct interaction between citizens by allowing them to communicate privately, rather than allowing them to observe a public signal of what others are doing.

2. Related Literature

Our paper builds on the global game literature (Carlsson and van Damme 1993 and Morris and Shin 1998), which has been applied to analyze issues in political economy. Among recent examples are Boix and Svolik (2010), Chassang and i Miquel (2010) and Edmond (2011). In our model, the dispersion in private information is crucial to generate different assessments on the informativeness of rumors. Information heterogeneity provides a ground for the study of communication between citizens, which is at the core of this paper. We choose to interpret our model in the context of revolution, but it can also help to understand similar coordination games, such as bank runs and currency attacks.

Our work is related to a small economics literature on rumors. Banerjee (1993) develops a model where rumors on investment opportunity are passed on from one agent to another, but recipients do not necessarily believe those rumors. In that model, the probability that someone hears a rumor is positively related to the number of people who have heard it. Bommel (2003) studies the effects of rumors on stock prices. In his model, informed investors with limited trading capacity profit from their private information, and they also spread rumors (i.e., give informative yet imprecise information to their followers) so that they profit from stock-price manipulation. Like their models, we also assume that a rumor heard by citizens cannot be verified, and that citizens use Bayesian updating in deciding whether to take action or not. Unlike their models, in which a rumor is passed on to each other sequentially, we provide a model in a static setting, in which a rumor is heard by citizens simultaneously. We focus on communication among citizens about the rumor rather than the transmission of the

[7] Angeletos and Werning (2006) is one of the few exceptions. They explicitly acknowledge the importance of direct interaction between agents in coordination models. In one extension of their model, agents are allowed to observe a public signal about the aggregate attack, which conveniently approximates the situation where agents could learn about actions of others. We capture the direct interaction between citizens by allowing them to discover other's personal judgment through one-to-one private communication.

rumor itself.

This paper also contributes to a growing literature on revolutions in economics. Edmond (2011) considers a coordination game where citizens' private information about the regime's strength is contaminated by the regime's propaganda. Citizens understand the signal-jamming technology used by the regime and form their beliefs accordingly. Our model differs in that private information is uncontaminated, but the public signal may be false and unrelated to the fundamental. Both Bueno de Mesquita (2010) and Angeletos and Werning (2006) study coordination games with two stages, where public signals arise from the first stage. In our model, the "attack stage" is preceded by a "communication stage," where a private message endogenously arises and enlarges citizens' information set. New development of this literature puts emphasis on uncertain payoffs from revolt (Bernhardt and Shadmehr 2010 and Bueno de Mesquita 2010). Given that our focus is the effect of rumors and communication, we assume that only the strength of the regime is uncertain.

In other fields of social sciences, there is no lack of discussions on rumors (e.g., Allport and Postman 1947) and revolutions (e.g., Goldstone 1994). However there are fewer studies on the relationship between these two. The idea seems to have been "up in the air" that rumors work as a tool to motivate citizens to participate in social movements such as riots, demonstrations and revolts, but the precise mechanisms through which the actions taken by citizens are related to the rumors they hear remain unspecified. Our model is a first step toward formalizing one such mechanism to explain explicitly how rumors affect citizens' beliefs, actions, and therefore equilibrium outcomes in revolutions.

3. A Model of Rumors and Talking about Rumors

3.1. Players and payoffs

Consider a society populated by a unit mass of ex ante identical citizens, indexed by $i \in [0,1]$, who play against another player, the regime. Citizen i chooses one of two actions: revolt ($a_i = 1$) or not revolt ($a_i = 0$). The aggregate mass of population that revolt is denoted A. Nature selects the strength of the regime, θ. The regime survives if and only if $\theta > A$; otherwise it is overthrown. A citizen's payoff depends both on whether the regime is overthrown and on whether the citizen chooses to revolt. A positive cost, $c \in (0,1)$, has to be paid if she revolts. If the regime is overthrown, citizens who revolt receive a benefit, $b = 1$, and those who do not participate receive

no benefit.[8] A citizen's net utility is therefore:

$$u(a_i, A, \theta) = \begin{cases} 1 - c, & \text{if } a_i = 1 \text{ and } A \geq \theta; \\ -c, & \text{if } a_i = 1 \text{ and } A < \theta; \\ 0, & \text{if } a_i = 0. \end{cases}$$

3.2. Information structure

Citizens are ex ante identical and have improper prior on θ. They become ex post heterogeneous after each of them observes a noisy private signal,

$$x_i = \theta + \varepsilon_i,$$

where the idiosyncratic noise $\varepsilon_i \sim \mathcal{N}(0, \sigma_x^2)$ is normally distributed and independent of θ, and is independently and identically distributed across i. This assumption captures the situation that citizens have diverse assessments of the regime's strength, before they hear any rumor and communicate.

Our next assumption is that all citizens hear a rumor, z, concerning the strength of the regime. The rumor is a public signal observed by all citizens. It could come from two alternative sources: either a source which offers an informative signal on the strength of the regime, or a source which only produces uninformative noise. Formally we model the random variable z as coming from a mixture distribution:

$$z \sim \begin{cases} I \sim \mathcal{N}(\theta, \sigma_z^2), & \text{with probability } \alpha; \\ U \sim \mathcal{N}(s, \sigma_U^2), & \text{with probability } 1 - \alpha; \end{cases}$$

where I indicates the informative source and U indicates the uninformative source. The parameters, s and σ_U^2, are the mean and variance of the uninformative distribution, respectively. The rumor comes from an informative source with prior probability α. When this is the case, z is normally distributed with mean θ and variance σ_z^2. We assume that α, s, σ_z, and σ_U are commonly known to all citizens.

[8] We abstract from free-riding issues, which has been carefully addressed by Edmond (2011) and Bernhardt and Shadmehr (2010). The benefits from regime change can be modeled as a public good that all citizens would enjoy. Edmond (2011) offers a general payoff structure to accommodate this concern. He shows that a condition can be derived such that citizens still have incentives to act against the regime, despite the free-riding incentives. To avoid being repetitive and keep the results sharp, we adopt a simpler payoff structure in this paper, which is a special case of his.

We maintain the following parameter restrictions throughout this paper:

$$\sigma_x < \sigma_z^2 \sqrt{2\pi}; \tag{1}$$

$$\sigma_U^2 > \sigma_x^2 + \sigma_z^2 \equiv \sigma_I^2. \tag{2}$$

The first restriction is standard. When $\alpha = 1$, the model reduces to the standard Morris-Shin model with public signal. Condition (1) is both sufficient and necessary for uniqueness of equilibrium in that model; see the discussion in Angeletos and Werning (2006). The second restriction captures the idea that uninformative noises exhibit greater variability than informative signals. Given that the rumor comes from informative sources, the variance of z conditional on θ is σ_z^2. Since θ has a variance of σ_x^2 to agents with private signal x_i, the unconditional variance of the informative source is $\sigma_I^2 \equiv \sigma_x^2 + \sigma_z^2$.

We stress that our specification of rumor as a mixture distribution is different from an informative public signal with low precision. According to the linear Bayesian updating formula, all agents would react to an informative public signal in the same way regardless of their private information. In our specification, however, citizens make an inference on how likely the rumor is informative. Since agents with different private signals have different views regarding what informative news would be like, they react to the same rumor differently. This mechanism plays a central role in our paper and cannot be replaced by modeling rumor simply as an informative public signal with a higher variance.

The parameter s can be interpreted as the "sentiment" of the public, which captures their perception of what uninformative messages would sound like. For example, if the public is used to receiving propaganda materials telling that the regime is strong, then they may expect a high value of s. Alternatively, sentiments may be driven by external events: news that would have seemed completely implausible (e.g., Mubarak leaving the country) could suddenly become plausible when similar news was reported and confirmed in a neighboring country. In this paper, we do not model where the public's sentiment s comes from, but we provide comparative statics analysis of how shifts in s affect equilibrium outcomes.

3.3. Communication

After citizens observed their private signals and the rumor, they are randomly paired up to communicate with each other. Specifically each citizen in a pair expresses her views on the likelihood that the rumor is drawn from an informative source, and hears her partner's views on the same matter. We assume that such communication is not frictionless: citizens can only convey to their partners whether they believe the rumor

is informative or not in a binary fashion. Let y_i represent the signal sent to citizen i by her partner j. The communication technology is characterized as follows.

$$y_i = \begin{cases} 1, & \text{if } \Pr[z \sim I | z, x_j] \geq \delta; \\ 0, & \text{if } \Pr[z \sim I | z, x_j] < \delta. \end{cases}$$

The threshold δ in this communication technology is common to all citizens and can be interpreted as a measure of their caution. Citizen j who sends the message $y_i = 1$ to citizen i can be interpreted as saying, "I believe the rumor is informative;" while the message $y_i = 0$ can be interpreted as "I don't believe it." A high value of δ means that citizens are unlikely to say they believe in the rumor unless they are sufficiently confident of their assessment. To rule out the possibility that agents will never say they believe in the rumor, we require that δ be not too high. Specifically we maintain the following assumption throughout the paper:

$$\delta < \frac{\alpha \sigma_I^{-1}}{\alpha \sigma_I^{-1} + (1-\alpha)\sigma_U^{-1}} \equiv \bar{\delta}. \tag{3}$$

A few comments on our assumption about the communication technology are in order. First, communication is non-strategic. Given that it is a game with a continuum of agents, and that each agent only communicates with one other agent, there is no incentive for citizens to strategically manipulate their partners' belief by lying.

Second, the binary message space captures the coarsening of information in the communication process. Citizens cannot fully express, explain and justify their own private assessment on the informativeness of the rumor to their partners. The transmission of information across individuals is beset with frictions that makes conveying the exact probability assessments difficult. Information coarsening and communication in a binary fashion can be a cost-effective way of exchanging casual information, and is commonly used in real life.

Third, we stress that talking about rumors is qualitatively different from other forms of communication. In section 6.1, we consider a model in which citizens exchange their private signals with their partners. Under that alternative technology, communication improves the precision of people's private information and makes them rely less on the publicly observed rumor. Under our maintained communication technology, on the other hand, we will show that communication reinforces the impact of the publicly observed rumor.

3.4. Decision rules

This model can be regarded as a two-stage game. In the communication stage, citizen i sends a message, $y_j \in \{0,1\}$, to her partner j based on the information set $\{z, x_i\}$, and receives a private message $y_i \in \{0,1\}$ from her partner. In the attack stage, citizen i chooses to revolt or not, given the post-communication information set $\{z, x_i, y_i\}$.

In the communication stage, citizens are heterogeneous in only one dimension: they observe diverse private signal x_i. Different people have different beliefs that the same rumor z is informative because they have different expectations of what informative news is like. They follow the standard Bayesian rule to update their beliefs:

$$\Pr[z \sim I|z, x_i] = \frac{\alpha\sigma_I^{-1}\phi\left(\sigma_I^{-1}(z - x_i)\right)}{\alpha\sigma_I^{-1}\phi\left(\sigma_I^{-1}(z - x_i)\right) + (1-\alpha)\sigma_U^{-1}\phi\left(\sigma_U^{-1}(z - s)\right)} \equiv w(z, x_i), \quad (4)$$

where ϕ is the standard normal density function.

The function $w(\cdot, x_i)$ is single-peaked in z, reaching a maximum at

$$z = x_i + \frac{\sigma_I^2}{\sigma_U^2 - \sigma_I^2}(x_i - s).$$

Thus rumors which take extremely high or low values are not believed by most citizens. This feature is a consequence of our assumption that the distribution of U has fatter tails than the distribution of I. We also note that $w(z, \cdot)$ is symmetric and single-peaked in x_i, reaching a maximum at $x_i = z$. This means that citizens whose private information are consistent with the rumor are more likely to believe that the rumor is informative.[9]

Since $w(z, \cdot)$ is single-peaked in x_i and reaches a maximum at $x_i = z$, there exists $\underline{x}(z)$ and $\overline{x}(z)$ such that

$$w(z, \underline{x}(z)) = w(z, \overline{x}(z)) = \delta,$$

provided $w(z, z) > \delta$. Assumption (3) on the upper bound of δ ensures that this condition is satisfied for any z. A citizen with private information x_i between $\underline{x}(z)$ and $\overline{x}(z)$ would have belief $w(z, x_i)$ greater than the threshold δ. The (non-strategic) decision rule in communication stage can therefore be characterized by:

$$y(z, x_i) = \begin{cases} 1, & \text{if } x_i \in [\underline{x}(z), \overline{x}(z)]\,; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

[9] See also Gentzkow and Shapiro (2006) and Suen (2010) for models which share the same feature.

Finally, note that the symmetry of $w(z,\cdot)$ about the point $x_i = z$ implies that $\underline{x}(z)$ and $\overline{x}(z)$ are centered about z. That is, we can write

$$\underline{x}(z) = z - \kappa(z) \quad \text{and} \quad \overline{x}(z) = z + \kappa(z),$$

for some $\kappa(z) > 0$.

After communication, the information set of citizen i is enlarged by the private message y_i sent by her partner j. Let $p(\theta|z, x_i, y_i)$ represent her posterior density function of θ. In the attack stage, she chooses a_i to maximize expected utility,

$$a(z, x_i, y_i) = \arg \max_{a_i \in \{0,1\}} \left\{ \int_{-\infty}^{\infty} u(a_i, A(\theta, z), \theta) p(\theta|z, x_i, y_i) \, \mathrm{d}\theta \right\} \tag{6}$$

The aggregate mass of attackers depends on the distribution of private signals x_i and communication messages y_i. Given state θ, x_i is normally distributed with mean θ and variance σ_x^2. Given the decision rule $y(z, x_i)$ specified in equation (5), the probability that citizen i receives the message $y_i = 1$ is equal to the probability that x_j lies within $[\underline{x}(z), \overline{x}(z)]$. Let $J(\theta, z)$ represent this probability conditional on θ and z. Then the aggregate mass of attackers is given by:

$$A(\theta, z) = \int_{-\infty}^{\infty} \left(J(\theta, z) a(z, x_i, 1) + (1 - J(\theta, z)) \, a(z, x_i, 0) \right) \frac{1}{\sigma_x} \phi \left(\frac{x_i - \theta}{\sigma_x} \right) \, \mathrm{d}x_i, \tag{7}$$

with

$$J(\theta, z) = \Phi \left(\frac{\overline{x}(z) - \theta}{\sigma_x} \right) - \Phi \left(\frac{\underline{x}(z) - \theta}{\sigma_x} \right), \tag{8}$$

where Φ is the standard normal distribution function.

3.5. Equilibrium

To summarize, the timing of events is the following. First, nature selects the strength of the regime. Then citizens receive private signals and hear a rumor. Afterwards, citizens are randomly matched in pairs and exchange their views on the informativeness of the rumor. Based on the updated information set, they choose to revolt or not. The regime survives or not, hinging on the measure of the mass of citizens who participate.

Definition 1. *An equilibrium is a set of posterior beliefs $p(\theta|z, x_i, y_i)$, a message sending decision $y(z, x_i)$, a revolt decision $a(z, x_i, y_j)$, and a mass of attackers $A(\theta, z)$, such that (5), (6) and (7) hold and the beliefs are derived from Bayes' rule.*

We focus on monotone equilibrium in which, for each realization of the rumor z, there is a threshold strength of the regime $\theta^*(z)$ such that the regime collapses if and

only if $\theta \leq \theta^*(z)$. Citizens adopt the following monotonic decision rule:

$$a(z, x_i, y_i) = \begin{cases} 1, & \text{if } x_i \leq x_I^*(z) \text{ and } y_i = 1, \\ & \text{or } x_i \leq x_U^*(z) \text{ and } y_i = 0, \\ 0, & \text{otherwise;} \end{cases}$$

where x_I^* and x_U^* are a pair of cut-off rules. The equilibrium ordering of x_U^* and x_I^* depends on the realization of z, and is elaborated in Section 5.

In a monotone equilibrium, the cut-off types must be indifferent between attacking and not attacking. Let $P(\cdot|z, x_i, y_i)$ be the cumulative distribution corresponding to the posterior density $p(\cdot|z, x_i, y_i)$. Then the indifference conditions can be written as:

$$P(\theta^*|z, x_I^*, 1) = c, \tag{9}$$

$$P(\theta^*|z, x_U^*, 0) = c. \tag{10}$$

Let $\hat{A}(\theta, \underline{x}(z), \overline{x}(z), x_I^*, x_U^*)$ be the mass of attackers when the state is θ and when citizens adopt the cut-off rules x_I^* and x_U^*. Then,

$$\hat{A}(\theta, x_I^*, x_U^*; z) = J(\theta, z)\Phi\left(\frac{x_I^* - \theta}{\sigma_x}\right) + (1 - J(\theta, z))\Phi\left(\frac{x_U^* - \theta}{\sigma_x}\right),$$

where the function J is given by (8). The threshold regime strength must satisfy

$$\hat{A}(\theta^*, x_I^*, x_U^*; z) = \theta^*. \tag{11}$$

A monotone equilibrium can be characterized by the triple (θ^*, x_I^*, x_U^*) that solves equations (9), (10), and (11).

4. Rumors without Communication

Our model departs from the standard global game model of Morris-Shin model in two respects: (1) the public signal may be uninformative; and (2) citizens can exchange messages concerning its informativeness . To highlight the effects of these two separate features, we discuss in this section a simplified model with feature (1) only but without communication. Such a model can be obtained as a special case of our model by setting $\delta = 0$, so that the everyone always sends the same message $y(z, x_i) = 1$ and communication becomes irrelevant. We refer to this special case (i.e., $\delta = 0$) as the "mute model." The "communication model" with both features (1) and (2) present (i.e., $\overline{\delta} > \delta > 0$) is left for Section 5.

The "mute model" nests two important benchmarks: $\alpha = 0$ and $\alpha = 1$. When

$\alpha = 0$, citizens believe that the rumor is completely uninformative. In this case, the value of z is irrelevant and the model reduces to a standard Morris-Shin model without public signal. We refer to this as the "pure noise model" and use $(\theta^*_{ms}, x^*_{ms})$ to denote the equilibrium threshold regime strength and cut-off agent type in this model.

The other important benchmark is the case of $\alpha = 1$. In this case, the rumor is known to be informative, and the model reduces to the standard Morris-Shin model with public signal. We refer to this case as the "public signal model" and use $(\theta^*_{ps}, x^*_{ps})$ to denote the equilibrium threshold state and cut-off type. Obviously in this case these equilibrium values depend on the realization of z.

The following result is standard. The uniqueness claim is established in, for example, Morris and Shin (2003) and Angeletos and Werning (2006). For completeness, we provide a proof of the remaining claims in the Appendix.

Proposition 1. *In the "pure noise model" (i.e., $\alpha = 0$), the equilibrium is $\theta^*_{ms} = 1 - c$ and $x^*_{ms} = (1 - c) - \sigma_x \Phi^{-1}(c)$. In the "public signal model" (i.e., $\alpha = 1$), there exists a unique equilibrium $\theta^*_{ps}(z)$ and $x^*_{ps}(z)$ such that:*

1. *$\theta^*_{ps}(z)$ and $x^*_{ps}(z)$ are decreasing in z;*
2. *(a) $\lim_{z\to\infty} \theta^*_{ps}(z) = 0$ and $\lim_{z\to\infty} x^*_{ps}(z) = -\infty$;*
   (b) $\lim_{z\to-\infty} \theta^*_{ps}(z) = 1$ and $\lim_{z\to-\infty} x^*_{ps}(z) = \infty$; and
3. *there exists a unique $\tilde{z}$ such that $\theta^*_{ps}(\tilde{z}) = \theta^*_{ms}$ and $x^*_{ps}(\tilde{z}) = x^*_{ms}$.*

In the "public signal model," citizen i's posterior mean of θ upon observing public signal z is

$$X_i = \beta x_i + (1 - \beta) z, \tag{12}$$

where $\beta = \sigma_z^2/(\sigma_z^2 + \sigma_x^2)$ (and the posterior variance is $\beta\sigma_x^2$). A lower value of z indicates that the regime is more fragile. Other things equal, this would result in more agents revolting against the regime (x^*_{ps} increases), making the regime more likely to collapse (θ^*_{ps} increases). In the limit, as z becomes very small, almost all citizens would revolt and the regime with type $\theta < 1$ would collapse almost surely (θ^*_{ps} goes to 1).

Part (1) and part (3) of Proposition 1 implies that $\theta^*_{ps}(z) > \theta^*_{ms}$ and $x^*_{ps}(z) > x^*_{ms}$ for all $z < \tilde{z}$. In what follows we say that a rumor is "against the regime" if $z < \tilde{z}$; and that it is "for the regime" if $z > \tilde{z}$.

In the "mute model," $\alpha \in (0, 1)$ and $\delta = 0$, so that citizens are skeptical of rumors but communication is ineffective. By Bayes' rule, the posterior belief about θ upon hearing a rumor z is a mixture of the posterior distribution in the "public signal model" and that in the "pure noise model," with weights given by the posterior belief that the

rumor is informative or not. In other words,

$$P(\theta|z, x_i) = w(z, x_i)\Phi\left(\frac{\theta - X_i}{\sqrt{\beta}\sigma_x}\right) + (1 - w(z, x_i))\Phi\left(\frac{\theta - x_i}{\sigma_x}\right), \tag{13}$$

where the weight function $w(z, x_i)$ is given by (4) and where the posterior mean X_i is given by (12) of the "public signal model." The crucial feature of this mixture distribution for θ is that the relative weights are not fixed (even though the mixture distribution for z have fixed weights α and $1 - \alpha$). Instead the weights depend on both the content of the rumor and on each citizen's private information.

Despite the dependence of the weights of the mixture distribution on private information, the cumulative distribution corresponding to posterior belief about θ is still stochastically increasing in x_i. To see this, let $h(x_i|\theta, z)$ be the density function of x_i given state θ and rumor z. For any $\theta_1 > \theta_0$, the likelihood ratio,

$$\frac{h(x_i|\theta_1, z)}{h(x_i|\theta_0, z)} = \frac{\phi(\sigma_x^{-1}(x_i - \theta_1))}{\phi(\sigma_x^{-1}(x_i - \theta_0))},$$

is increasing in x_i. This monotone likelihood ratio property implies that the posterior distribution, $P(\theta|z, x_i)$, is decreasing in x_i (Milgrom 1981). Thus, for any threshold value of the regime strength θ, the expected payoff from revolt is decreasing in the value of the private information x_i. This justifies our focus on monotone equilibrium.

Let (θ_m^*, x_m^*) represent the equilibrium threshold regime strength and cut-off agent type in the "mute model." The cut-off type must be indifferent between revolt and not revolt:

$$P(\theta_m^*|z, x_m^*) = c; \tag{14}$$

and the mass of attackers in state θ_m^* must be equal to the threshold regime strength:

$$\hat{A}(\theta_m^*, x_m^*) = \Phi\left(\frac{x_m^* - \theta_m^*}{\sigma_x}\right) = \theta_m^*. \tag{15}$$

These equilibrium values depend on the realization of the rumor. The following result characterizes the equilibrium $\theta_m^*(z)$ and compares it with the benchmark "pure noise model" and "public signal model."

Proposition 2. *In the "mute model" (i.e., $\alpha \in (0, 1)$ and $\delta = 0$), the equilibrium threshold regime strength $\theta_m^*(z)$ satisfies:*

1. $\theta_m^*(\tilde{z}) = \theta_{ps}^*(\tilde{z}) = \theta_{ms}^*$;
2. *(a) if the rumor is against the regime (i.e., $z < \tilde{z}$), then $\theta_m^*(z) \in (\theta_{ms}^*, \theta_{ps}^*(z))$;*
   (b) if the rumor is for the regime (i.e., $z > \tilde{z}$), then $\theta_m^*(z) \in (\theta_{ps}^*(z), \theta_{ms}^*)$;

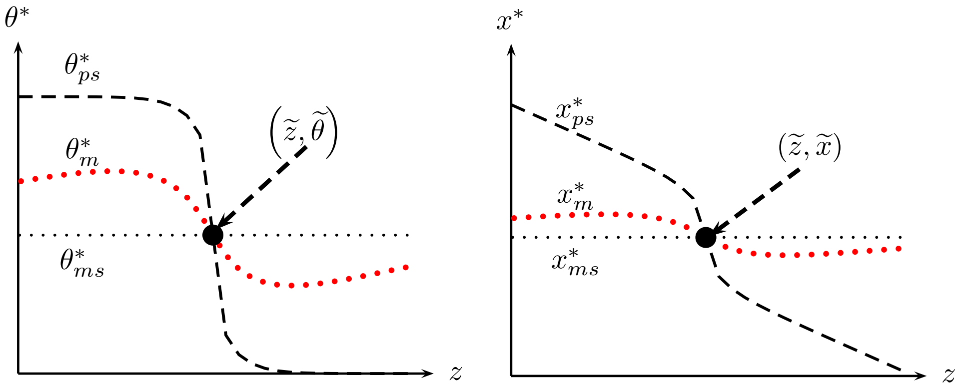
3. $\lim_{z\to-\infty} \theta^*_m(z) = \lim_{z\to\infty} \theta^*_m(z) = \theta^*_{ms}$; *and*

4. $\theta^*_m(z)$ *is increasing then decreasing then increasing.*

Figure 1(a) illustrates the properties of the equilibrium threshold strength $\theta^*_m(z)$ described in this proposition.[10] Figure 1(b) shows the qualitative properties of the equilibrium cut-off agent type $x^*_m(z)$ in the "mute model." These properties are similar to those of $\theta^*_m(z)$: $x^*_m(z)$ is increasing then decreasing then increasing, with $\lim_{z\to\pm\infty} x^*_m(z) = x^*_{ms}$. This follows from the attack equation (15), which can be written as:

$$x^*_m(z) = \theta^*_m(z) + \sigma_x \Phi^{-1}(\theta^*_m(z)).$$

Assumption (1) guarantees that $x^*_m(z)$ is increasing in $\theta^*_m(z)$.



(a) threshold regime strength $\theta^*(z)$ (b) cut-off agent type $x^*(z)$

Figure 1. *Comparing the "mute model" with benchmark "pure noise model" and "public signal model"*

Interestingly, the equilibrium pair $(\theta^*_{ms}, x^*_{ms})$ of the "pure noise model" also solves the "mute model" when the rumor z is neither "for the regime" nor "against the regime." Recall that in the "pure noise model" an agent with private information x^*_{ms} believes that the regime will collapse with probability c given threshold regime strength θ^*_{ms}. Given the same threshold regime strength θ^*_{ms}, this agent also believes that the regime will collapse with the same probability c in the "public signal model" if the public signal is $\tilde{z}$. Thus, for $z = \tilde{z}$, it does not matter for this agent whether the rumor is informative or not. If the threshold regime strength is θ^*_{ms}, she believes that the regime will collapse with probability c. Thus the indifference condition (14) is satisfied. Given that the cut-off agent type is x^*_{ms} at $z = \tilde{z}$, the threshold θ^*_{ms} satisfies the attack equation (15). Thus $(\theta^*_{ms}, x^*_{ms})$ is an equilibrium of the "mute model" at $z = \tilde{z}$.

In this model, citizens are skeptical of the rumor, i.e., they take into account the

[10] Unless otherwise specified, we use the following set of parameters as benchmark to compute numerical examples and illustrate: $c = 0.5$, $s = 0.5$, $\alpha = 0.5$, $\delta = 0.5$, $\sigma^2_U = 1$, $\sigma^2_z = 0.5$, and $\sigma^2_x = 0.4$.

possibility that the rumor could just provide a noise. The degree of skepticism is characterized by the weight given to the rumor being informative, $w(\tilde{z}, \tilde{x}) \in (0, 1)$. The effect of skepticism manifests itself in the fact that the equilibrium regime strength threshold is less sensitive to changes in z around $\tilde{z}$ than it is in the "public signal model" and more responsive than in the "pure noise model." Specifically,

$$0 = \frac{\mathrm{d}\theta^*_{ms}}{\mathrm{d}z} > \frac{\mathrm{d}\theta^*_{m}(\tilde{z})}{\mathrm{d}z} > \frac{\mathrm{d}\theta^*_{ps}(\tilde{z})}{\mathrm{d}z}, \tag{16}$$

which is implied by part (2) of Proposition 2.

To understand why (16) holds, let $\hat{x}(\theta, z)$ be defined as the cut-off agent type that solves the indifference condition at threshold regime strength θ and rumor realization z. In the "mute model," as well as in the benchmark models, the response of equilibrium threshold to a change in the realization of the rumor can be calculated by,

$$\frac{\mathrm{d}\theta^*(\tilde{z})}{\mathrm{d}z} = \frac{\frac{1}{\sigma_x}\phi(\cdot)\frac{\partial \hat{x}}{\partial z}}{1 - \frac{1}{\sigma_x}\phi(\cdot)\left(-1 + \frac{\partial \hat{x}}{\partial \theta}\right)}, \tag{17}$$

where $\phi(\cdot)$ is evaluated at the equilibrium value of $(x^* - \theta^*)/\sigma_x$ of the relevant model. The effect of rumor on the equilibrium regime strength threshold can be decomposed into two components. At $z = \tilde{z}$, a lower value of z means that the regime is relatively weaker. In order to keep the expected payoff from attacking fixed, the cut-off agent type $\hat{x}$ must have a higher value of private information. Thus $\partial \hat{x}/\partial z < 0$, and the numerator of (17) is negative. As the threshold for regime survival rises, the payoff from attacking also rises, which raises $\hat{x}$ further, because $\partial \hat{x}/\partial \theta > 0$. This multiplier effect can be seen in the denominator of (17).

When $z = \tilde{z}$, the equilibrium values of θ^* and x^* are the same across the three models. Therefore, in equation (17), the term $\phi(\cdot)$ does not depend on the model being considered. The critical factors in determining the magnitude of $\mathrm{d}\theta^*/\mathrm{d}z$ are $\partial \hat{x}/\partial z$ and $\partial \hat{x}/\partial \theta$, which determine the magnitude of the direct and multiplier effects. At $z = \tilde{z}$, we have:

$$\frac{\partial \hat{x}_{ms}}{\partial \theta} = 1 < \frac{\partial \hat{x}_{m}}{\partial \theta} = \frac{w + (1-w)\sqrt{\beta}}{w\beta + (1-w)\sqrt{\beta}} < \frac{\partial \hat{x}_{ps}}{\partial \theta} = \frac{1}{\beta}. \tag{18}$$

Further, at $z = \tilde{z}$, $\partial \hat{x}/\partial z + \partial \hat{x}/\partial \theta = 1$ for all the three models. Therefore, (18) implies:

$$0 = \frac{\partial \hat{x}_{ms}}{\partial z} > \frac{\partial \hat{x}_{m}}{\partial z} > \frac{\partial \hat{x}_{ps}}{\partial z}.$$

The comparison of $\mathrm{d}\theta^*_m(\tilde{z})/\mathrm{d}z$ with the counterparts in the benchmark models given in (16) then follows immediately.

The comparison in (16) shows that, at the point $z = \tilde{z}$ the sensitivity of equilibrium outcome to the rumor in the "mute model" is between that in the "pure noise model" and in the "public signal model." More generally, part (2) of Proposition 2 states that the value of $\theta^*_m(z)$ is always between θ^*_{ms} and $\theta^*_{ps}(z)$ for any z. This, however, does not mean that the "mute model" is simply a "public signal model" with a low precision.

Part (4) of Proposition 2 states that $\theta^*_m(z)$ is non-monotone, while Proposition 1 states that $\theta^*_{ps}(z)$ is monotonic decreasing. Note that

$$\frac{\partial P(\theta^*_m|z, x_i)}{\partial z} = -w\frac{\beta}{\sqrt{\beta}\sigma_x}\phi_I + (\Phi_I - \Phi_U)\frac{\partial w}{\partial z},$$

where the subscripts I and U means that the functions are evaluated at the points $(\theta^*_m - X_i)/(\sqrt{\beta}\sigma_x)$ and $(\theta^*_m - x_i)/\sigma_x$, respectively. If the rumor is informative, an increase in z is indication that the regime is strong, which lowers the probability that the regime will collapse. Hence the first term is negative. However, since $w(z, x_i)$ is single-peaked in z, $\partial w/\partial z < 0$ for z sufficiently large. When the rumor is for the regime, the probability that the regime will collapse is lower if the rumor is informative than if the rumor is uninformative; that is, $\Phi_I < \Phi_U$. Therefore the second term is positive for z sufficiently large. When the second term dominates the first term, an increase in the value of z actually increases agents' assessment of the likelihood that the regime would collapse. As a result, the marginal agent who is indifferent between attacking and not attacking must have a higher private information about the strength of the regime. In other words, $\partial\hat{x}_m/\partial z > 0$ for z large, which implies $\mathrm{d}\theta^*_m(z)/\mathrm{d}z > 0$ by equation (17). When the value of z is intermediate, however, the first effect dominates, which means that a higher value of z is a signal that the regime is stronger. In this case, $\partial\hat{x}_m/\partial z < 0$, which implies $\mathrm{d}\theta^*_m(z)/\mathrm{d}z < 0$. This explains the non-monotonicity of $\theta^*_m(z)$.

The non-monotonicity result also explains why the limit behavior of the "mute model" is qualitatively different from that of the "public signal model." When the rumor is extreme, i.e., $z \to -\infty$ or $+\infty$, the probability that it comes from an informative source goes to 0. As a result, people disregard the rumor and the equilibrium is identical to that in the "pure noise" case: $\theta^*_m(z)$ goes to θ^*_{ms} and $x^*_m(z)$ goes to x^*_{ms}. In contrast, in the "public signal model", the equilibrium threshold monotonically deceases from 1 to 0 as z goes from $-\infty$ to ∞.

5. Rumors with Communication

In this section, we focus on the "communication model," with $\delta \in (0, \overline{\delta})$ and $\alpha \in (0, 1)$. In this model, citizens are still skeptical of the rumor they hear, but are allowed to

communicate with one another about the informativeness of the rumor. We show that communication can overcome the effect of skepticism and spark a sharp reaction to the rumor among citizens when the rumor is not extreme. More surprisingly, a rumor against the regime could be more effective in mobilization than a negative trustworthy news that all citizens fully believe to be informative.

5.1. Equilibrium Properties

By Bayes' rule, agent i who receives the message $y_i = 1$ from her peer revises her belief about the state to:

$$P(\theta|z, x_i, 1) = \frac{\int_{-\infty}^{\theta} J(t,z)p(t|z,x_i)\,\mathrm{d}t}{\int_{-\infty}^{\infty} J(t,z)p(t|z,x_i)\,\mathrm{d}t}.$$

where $p(\cdot|z, x_i)$ is the density associated with the belief $P(\cdot|z, x_i)$ in the "mute model" given by equation (13), and where $J(t,z)$ is the probability that a randomly selected agent would send a message to his peer that confirms the rumor z in state t, as given by equation (8). Note that the denominator is equal to $\Pr[y_i = 1|z, x_i]$, that is, the probability that agent i expects to receive $y_i = 1$ from his peer. The term $J(t,z)$ is increasing than decreasing in t, with a peak at $t = z$. Therefore, $J(t,z)/\Pr[y_i = 1|z, x_i]$ is higher than 1 when the state t is close to z, and is lower than 1 when t is far away from z. In other words, upon receiving the message $y_i = 1$, the posterior density becomes more concentrated around the rumor z than its counterpart in the "mute model." Similarly, agent i who receives the message $y_i = 0$ revises her belief to:

$$P(\theta|z, x_i, 0) = \frac{\int_{-\infty}^{\theta} (1 - J(t,z))p(t|z,x_i)\,\mathrm{d}t}{\int_{-\infty}^{\infty} (1 - J(t,z))p(t|z,x_i)\,\mathrm{d}t}.$$

Let $\hat{x}_I(\theta, z)$ be the value of x_i that solves $P(\theta|z, x_i, 1) = c$. That is, an agent with private information $\hat{x}_I$ is indifferent between attacking or not attacking if the threshold regime strength is θ, the rumor realization is z, and the peer's message is that she "believes the rumor." Similarly, let $\hat{x}_U(\theta, z)$ be the value of x_i that solves $P(\theta|z, x_i, 0) = c$. In a monotone equilibrium, there are three possible cases for the ordering of the cut-off types: $\hat{x}_I > \hat{x}_U$, $\hat{x}_I < \hat{x}_U$, or $\hat{x}_I = \hat{x}_U$. Suppose $\hat{x}_I > \hat{x}_U$. Then, citizens with private information $x_i < \hat{x}_U$ attack the regime regardless of the message they receive. We label this group of citizens *revolutionaries*. Similarly, citizens with private information $x_i > \hat{x}_I$ would not revolt regardless. This group of citizens is labeled *bystanders*. Citizens with $x_i \in [\hat{x}_U, \hat{x}_I]$ choose to revolt if they receive $y_i = 1$ from their partners, and not to revolt otherwise. Essentially, this group of citizens' revolt decisions are influenced by their partners' assessments of the rumor. We call this group the *swing population*. See Figure 2. Similarly, if $\hat{x}_I(z) < \hat{x}_U(z)$, revolutionaries are

citizens with private information $x_i < \hat{x}_I$, bystanders are those with $x_i > \hat{x}_U$, and the rest of citizens constitute the swing population. If $\hat{x}_I(z) = \hat{x}_U(z)$, the mass of swing population in society is zero.

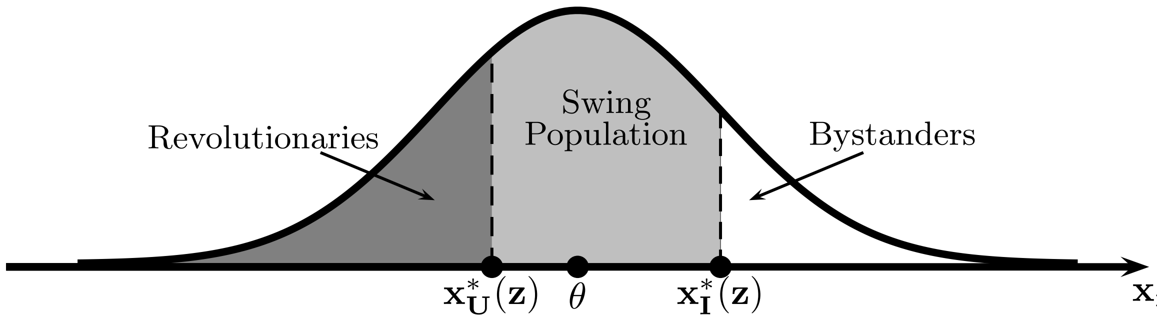


Figure 2. *An example of population distribution: revolutionaries, swing population and bystanders, when $x_U^* < x_I^*$.*

Proposition 3. *In the "communication model" (i.e., $\alpha \in (0,1)$ and $\delta \in (0,\bar{\delta})$), the equilibrium triple (θ^*, x_I^*, x_U^*) has the following properties:*

1. $\lim_{z\to\pm\infty} \theta^*(z) = \theta^*_{ms}$, $\lim_{z\to\pm\infty} x_U^*(z) = x^*_{ms}$, *and* $\lim_{z\to\pm\infty} x_I^*(z) = \mp\infty$.
2. *There exists a z' such that $x_I^*(z') = x_U^*(z') = x_m^*(z')$ and $\theta^*(z') = \theta^*_m(z')$.*
3. *If $c = 0.5$, then $z' = \tilde{z}$ and $\theta^*(z') = \theta^*_m(z') = \theta^*_{ps}(z') = \theta^*_{ms}$.*

The limit behavior of the "communication model" is summarized in part (1). When the rumor takes on extreme values, almost everyone says "I don't believe it" (because the limit of $J(\theta, z)$ is 0 for any finite θ). The skepticism of an agent (captured by the fact that the limit of $w(z, x_i)$ is 0) is reinforced by the skepticism of her peer (captured by the fact that she receives $y_i = 0$ almost surely). This explains why θ^* and x_U^* are the same as those in "pure noise model."

The limit behavior of x_I^*, however, is qualitatively different. If the rumor indicates that the regime is very weak, an agent i is extremely unlikely to receive a confirmatory message from her peer. But in the unlikely event that she does, such an extreme event will overcome her initial skepticism toward the rumor to the extent that, for any finite value of x_i, she becomes almost sure that the regime will collapse. Therefore, in equilibrium, the cutoff type $x_I^*(z)$ must increase without bound as z becomes very low, meaning that almost anyone who hears a confirmatory message about a rumor that the regime is extremely weak will revolt.

The limit properties in part (1) of Proposition 3 immediately imply that there exists z' such that $x_I^*(z') = x_U^*(z') = x'$. At such a z', the "communication model" is observationally equivalent to the "mute model": the mass of swing population is zero and

society is populated only by revolutionaries (citizens whose $x_i < x'$) and bystanders (citizens whose $x_i \geq x'$).

Interestingly, we find that such a z' coincides with $\tilde{z}$ in the case where $c = 0.5$. It means that rumors for which cut-off type citizens do not care about what their partners say ($y = 1$ or $y = 0$) are also rumors for which they do not care about from which source the rumor is drawn ($z \sim I$ or $z \sim U$). This special case offers analytical convenience and tractability, since equilibrium thresholds are the same across all the four models when $z' = \tilde{z}$, that is $\theta^*(\tilde{z}) = \theta^*_m(\tilde{z}) = \theta^*_{ps}(\tilde{z}) = \theta^*_{ms}$. We study the role played by communication in the "communication model" by letting z deviate from z' and compare its results with the other three benchmarks. Therefore, in the remaining analysis, we focus on this special case.

The equilibrium ordering of $x^*_I(z)$ and $x^*_U(z)$ depends on z. Specifically, $x^*_I(z) > x^*_U(z)$ if $z < z'$, and $x^*_I(z) < x^*_U(z)$ if $z > z'$. When $z < z'$, a citizen from the swing population (i.e., $x_i \in [x^*_U(z), x^*_I(z)]$) attacks the regime when he receives a message that confirms the rumor and does not attack otherwise. Intuitively, if the rumor says the regime is weak, i.e, $z < z'$, confirmatory messages from peers encourage the swing population to attack. When $z > z'$, the opposite is true.

To illustrate, we plot expected net payoffs $P(\theta^*(z)|z, x_i, y_i) - c$ against x_i for $y_i = 1$ and $y_i = 0$ in Figure 3. Figure 3(a) presents the case where $z = z'$. The net payoff crosses zero at $x_i = x'$ for both the case of $y_i = 1$ and the case of $y_i = 0$.[11] Therefore $x^*_I(z') = x^*_U(z') = x'$. Intuitively, the cut-off type citizen with $x_i = x'$ is indifferent between $y_i = 1$ and $y_i = 0$.

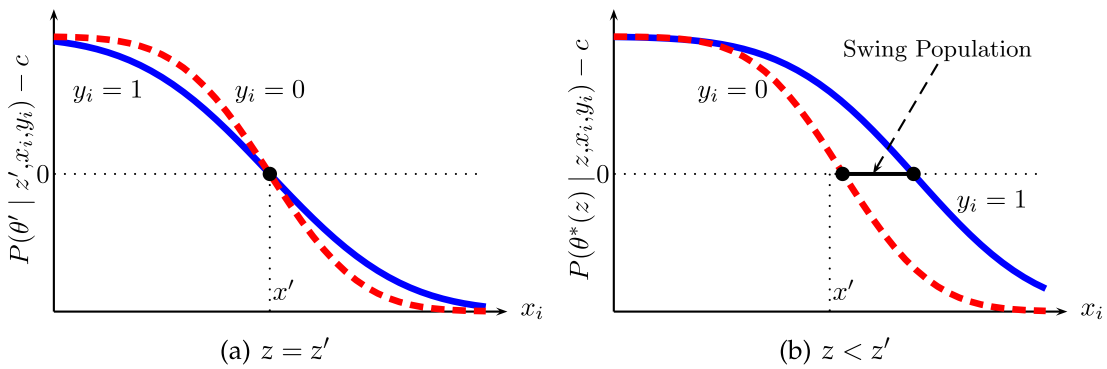


Figure 3. *Net payoff from attacking as a function of private information x_i*

Figure 3(b) presents the case where z is slightly below z'. A lower value of z is a

[11]The curve with $y_i = 1$ is flatter than the one with $y_i = 0$. Citizens with low private signals, $x_i < x'$, consider that z' indicates the regime is relatively stronger than they privately believe. Therefore, their expected net payoff is also lower than when they receive a message that confirms the rumor rather than one that casts doubt on the rumor. In other words, $P(\theta'|z', x_i, 1) < P(\theta'|z', x_i, 0)$ when $x_i < x'$. Similarly, $P(\theta'|z', x_i, 1) > P(\theta'|z', x_i, 0)$ when $x_i > x'$.

rumor against the regime, which, as shown in the subsection that follows, shifts up x_I^* and results in $x_I^* > x_U^*$. The gap between x_I^* and x_U^* shown in the figure indicates the swing population. Anyone from the swing population finds that the expected net payoff of attacking is higher than 0 when $y_i = 1$ and is lower than 0 when $y_i = 0$.

5.2. The Effects of Communication

In both the "communication model" and the "mute model," at $z = z'$, cut-off types with private information x' believe that the probability of successful revolution is c; that is,

$$P(\theta'|z', x', 1) = P(\theta'|z', x') = P(\theta'|z', x', 0) = c.$$

When the threshold regime strength θ changes, the cut-off type $\hat{x}(\theta, z)$ who is indifferent between attacking and not attacking also changes. Interestingly, the sensitivity of the cut-off type to changes in the threshold regime strength depends on the content of the communication, with individuals who receive $y_i = 1$ being more sensitive than individuals who receive $y_i = 0$, and with individuals in the "mute model" being in between. We label this ordering the "communication effect." Specifically Lemma 2 in the Technical Appendix establishes that, at the point (θ', z'),

$$\frac{\partial \hat{x}_I}{\partial \theta} = \frac{-p(\theta'|z', x')}{\int_{-\infty}^{\theta'} \frac{J(\theta, z')}{J(\theta', z')} \frac{\partial p}{\partial x} \, \mathrm{d}\theta} > \frac{\partial \hat{x}_m}{\partial \theta} = \frac{-p(\theta'|z', x')}{\int_{-\infty}^{\theta'} \frac{\partial p}{\partial x} \, \mathrm{d}\theta} > \frac{\partial \hat{x}_U}{\partial \theta} = \frac{-p(\theta'|z', x')}{\int_{-\infty}^{\theta'} \frac{1-J(\theta, z')}{1-J(\theta', z')} \frac{\partial p}{\partial x} \, \mathrm{d}\theta} > 0. \tag{19}$$

Furthermore, since $\partial \hat{x}/\partial z = 1 - \partial \hat{x}/\partial \theta$ at the point (θ', z') (see Lemma 3 in Technical Appendix), (19) also implies

$$\frac{\partial \hat{x}_I}{\partial z} < \frac{\partial \hat{x}_m}{\partial z} < \frac{\partial \hat{x}_U}{\partial z}. \tag{20}$$

Given inequality (19), (20) and the fact that $\mathrm{d}x^*/\mathrm{d}z = \partial \hat{x}/\partial z + \partial \hat{x}/\partial \theta \cdot \mathrm{d}\theta^*/\mathrm{d}z$, it is also straightforward to show that, at $z = z'$,

$$\frac{\mathrm{d}x_I^*}{\mathrm{d}z} < \frac{\mathrm{d}x_U^*}{\mathrm{d}z}.$$

This implies that $x_I^*(z) > x_U^*(z)$ for z slightly below z', and $x_I^*(z) < x_U^*(z)$ for z slightly above z'.

To understand equation (19), we first note that $z' = \theta'$ in this case. With a confirmatory message from her peer, a citizen believe that the true state θ is more likely to be close to z' and less likely to be far away from z'. He therefore assigns a smaller "weight" to those possible realizations of θ that are further away from z'. Specifically, conditional on receiving the message $y_i = 1$, the relative likelihood of any two states, say θ and θ', is given by the unconditional relative likelihood times the likelihood ratio

for $x_j \in [\underline{x}(z'), \overline{x}(z')]$:

$$\frac{p(\theta|z',x',1)}{p(\theta'|z',x',1)} = \frac{p(\theta|z',x')}{p(\theta'|z',x')}\frac{J(\theta,z')}{J(\theta',z')}.$$

Since $J(\theta,z')$ reaches a peak at $\theta = \theta'$, the likelihood ratio $J(\theta,z')/J(\theta',z')$ is smaller than 1. Similarly, the message $y_i = 0$ gives rise to the opposite weighting adjustment: $(1-J(\theta,z'))/(1-J(\theta',z')) \geq 1$. Moreover, for each possible realization $\theta < \theta'$, we have $\partial p(\theta|z',x')/\partial x < 0$. Therefore, the following inequality is established,

$$0 > \int_{-\infty}^{\theta'} \frac{J(\theta,z')}{J(\theta',z')}\frac{\partial p}{\partial x}\,\mathrm{d}\theta > \int_{-\infty}^{\theta'} \frac{\partial p}{\partial x}\,\mathrm{d}\theta > \int_{-\infty}^{\theta'} \frac{1-J(\theta,z')}{1-J(\theta',z')}\frac{\partial p}{\partial x}\,\mathrm{d}\theta,$$

which immediately leads to (19).

Intuitively, compared to "mute model," a confirmatory message ($y_i = 1$) leads to a more concentrated posterior density for states around $\theta = z$. To see this, note that there is an interval containing z such that if θ belongs to that interval then $J(\theta,z)/\Pr[y_i = 1|z,x_i] > 1$, and if θ is outside that interval then $J(\theta,z)/\Pr[y_i = 1|z,x_i] < 1$. As a result, this posterior distribution is less responsive to changes in private information than its counterpart in the "mute model." See Figure 4 for an illustration. The same amount of change in x shifts the density function $p(\cdot|z,x_i)$ to the right by a greater amount than it does to the density function $p(\cdot|z,x_i,1)$. In other words, a citizen who receives $y_i = 1$ relies less on his own private signal. That explains why $\hat{x}_I$ has to change by a larger amount than $\hat{x}_m$ to balance the indifference condition, in response to changes in regime strength threshold. The opposite is true for a citizen with message $y_i = 0$, which leads to a posterior density function $p(\cdot|z,x_i,0)$, that is less concentrated around z and is more responsive to changes in x_i than is the density function $p(\cdot|z,x_i)$.

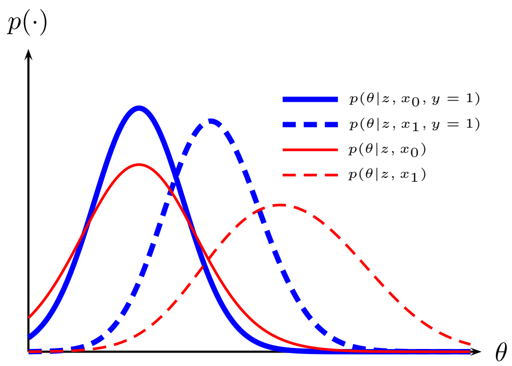


Figure 4. *Posterior density in the "communication model" and the "mute model." When $x_0 = z$, both densities are symmetric around z. When $x_1 > z$, density distribution in the "mute model" is shifted to the right by a greater amount than density distribution in "communication model," for the case $y = 1$.*

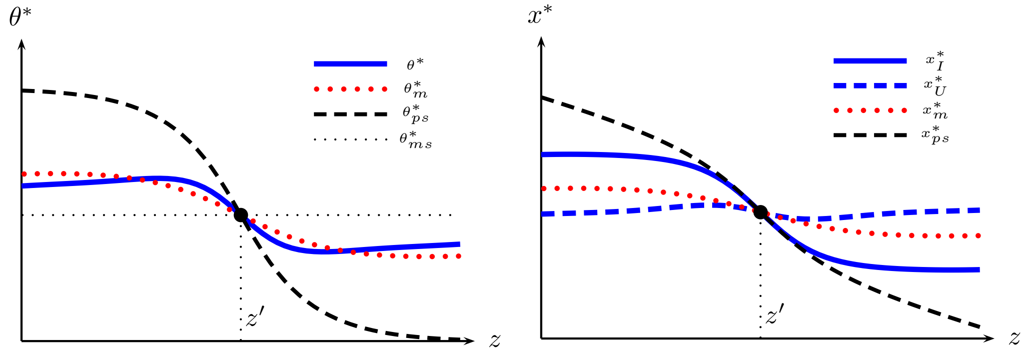
The following proposition establishes another key result: when the rumor is close

to neutral, the equilibrium threshold regime strength is more sensitive to the rumor under the "communication model" than under the "mute model." In other words, when the rumor z is against the regime and is sufficiently close to z', a regime with strength $\theta \in (\theta_m^*(z), \theta^*(z))$ would collapse if communication among citizens is allowed, but would survive if it is not.

Proposition 4. *At $z = z' = \tilde{z}$,*

$$\frac{\mathrm{d}\theta^*(z')}{\mathrm{d}z} < \frac{\mathrm{d}\theta_m^*(z')}{\mathrm{d}z} < 0, \tag{21}$$

Figures 5(a) and 5(b) present the equilibrium thresholds θ^* and the equilibrium cutoff rules x^* for the four models. Around z', the equilibrium θ^* in the "communication model" is steeper than that in the "mute model," as is described in Proposition 4.



(a) θ^*, θ_m^*, θ_{ps}^*, and θ_{ms}^* (b) x_I^*, x_U^*, x_m^* and x_{ps}^*

Figure 5. *Equilibrium threshold regime strength and equilibrium cutoff rules in the "communication model" in comparison to other models*

The size of $\mathrm{d}\theta^*/\mathrm{d}z$ is determined by the magnitude of the direct and multiplier effects. Using the fact that $\hat{x}_I(\theta', z') = \hat{x}_U(\theta', z')$, we can differentiate the attack equation (11) to obtain:

$$\frac{\mathrm{d}\theta^*(z')}{\mathrm{d}z} = \frac{\frac{1}{\sigma_x}\phi(\cdot)\left(J\frac{\partial \hat{x}_I}{\partial z} + (1-J)\frac{\partial \hat{x}_U}{\partial z}\right)}{1 - \frac{1}{\sigma_x}\phi(\cdot)\left(-1 + J\frac{\partial \hat{x}_I}{\partial \theta} + (1-J)\frac{\partial \hat{x}_U}{\partial \theta}\right)}, \tag{22}$$

where $\phi(\cdot)$ is evaluated at the point $(x' - \theta')/\sigma_x$. Comparing (22) with its counterpart (17) for the "mute model," we obtain Proposition 4 by showing in the proof that, at the point $\theta = \theta'$ and $z = z'$,

$$J\frac{\partial \hat{x}_I}{\partial \theta} + (1-J)\frac{\partial \hat{x}_U}{\partial \theta} > \frac{\partial \hat{x}_m}{\partial \theta} > 1; \tag{23}$$

$$J\frac{\partial \hat{x}_I}{\partial z} + (1-J)\frac{\partial \hat{x}_U}{\partial z} < \frac{\partial \hat{x}_m}{\partial z} < 0. \tag{24}$$

The direct effect of an increase in z on the mass of attackers is the weighted average of $\partial\hat{x}_I/\partial z$ and $\partial\hat{x}_U/\partial z$, with the weights being J and $1-J$, respectively. By (24), this effect is larger in magnitude with communication than without. The feedback effect of an increase in threshold regime strength on the mass of attackers is determined by the weighted average of $\partial\hat{x}_I/\partial\theta$ and $\partial\hat{x}_U/\partial\theta$. By (23), this effect is also larger in magnitude with communication than without. These two effects combine to give a greater sensitivity of the equilibrium regime survival threshold to the rumor.

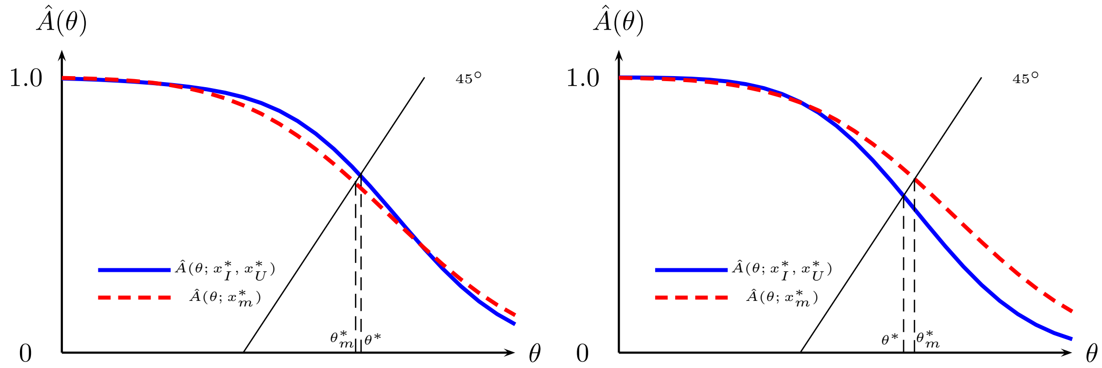
When the true state of nature is $\theta = \theta'$, the mass of attackers is equal to the equilibrium threshold regime strength. Therefore, Proposition 4 also implies that the mass of attackers in state $\theta = \theta'$ is more responsive to the rumor in the "communication model" than in the "mute model." A slight decrease in z from z' leads to an increase in the size of swing population by $\sigma_x^{-1}\phi(\cdot)(\mathrm{d}x_I^*/\mathrm{d}z - \mathrm{d}x_U^*/\mathrm{d}z)$. A fraction J of the swing population would receive a confirmatory message from their peers and attack the regime. The change in the mass of revolutionaries is $\sigma_x^{-1}\phi(\cdot)\mathrm{d}x_U^*/\mathrm{d}z$. In the "mute model", the increase in the mass of attackers is $\sigma_x^{-1}\phi(\cdot)\mathrm{d}x_m^*/\mathrm{d}z$. At $z = z'$, we must have

$$\frac{1}{\sigma_x}\phi(\cdot)\left[J(\theta', z')\left(\frac{\mathrm{d}x_I^*}{\mathrm{d}z} - \frac{\mathrm{d}x_U^*}{\mathrm{d}z}\right) + \frac{\mathrm{d}x_U^*}{\mathrm{d}z}\right] < \frac{1}{\sigma_x}\phi(\cdot)\frac{\mathrm{d}x_m^*}{\mathrm{d}z} < 0.$$

To illustrate, we plot the mass of attackers $\hat{A}(\theta)$ against the realization of regime strength θ, holding the cut-off rules constant. Equilibrium threshold is given by the intersection of $\hat{A}(\theta)$ and the 45-degree line. For values of z slightly below z', Figure 6(a) shows that a larger fraction of attackers are mobilized than that in the "mute model" when the regime strength θ is near z. Note that communication does not increase the mass of attackers at all states. Figure 6(a) shows that communication actually lowers the mass of attackers when θ is far from z.

Communication among citizens reveals extra information regarding the rumor and the underlying unknown state. Because the function $J(\theta, z)$ peaks at $\theta = z$, more citizens will hear a confirmatory message when the true state is near what the rumor suggests. If the rumor suggests that the regime is weak, this mechanism causes more people to attack when the regime is indeed weak and close to what the rumor indicates. This explains why $\hat{A}(\theta')$ is higher in the "communication model" than that in the "mute model," or why $\theta^*(z) > \theta_m^*(z)$ for z slightly lower than z'.

For the same reason, the ordering of $\theta^*(z)$ and θ_m^* may be reversed for z much below z'. This is illustrated in Figure 6(b). When the true regime strength θ is much higher than what the rumor z suggests, many citizens will hear that their peers do not believe the rumor. In this case, communication coordinates citizens not to attack. Figure 6(b) shows that the mass of attackers falls relative to that in the "mute model"



(a) $\hat{A}(\theta; x_I^*, x_U^*)$ and $\hat{A}(\theta; x_m^*)$, when $z < z'$ (b) $\hat{A}(\theta; x_I^*, x_U^*)$ and $\hat{A}(\theta; x_m^*)$, when $z << z'$

Figure 6. *Mass of attackers $\hat{A}(\theta; \cdot)$ in different states θ in the "mute model" and the "communication model"*

when regime is indeed strong.

5.3. Rumors vs. Trustworthy News

In section 4, we have shown that, due to citizens' skepticism toward rumors, the magnitude of the response of θ_m^* to a change in z is smaller in the "mute model" than that in the "public signal model" at $z = \tilde{z}$. In the previous subsection, we establish that the effect of skepticism can be, to a certain extent, reduced by communication, so that θ^* is more sensitive than θ_m^* is to changes in z around z'. Interestingly, when the precision of the private signal is reasonably high, the effect of communication can be so large that θ^* in the "communication model" is more sensitive to the rumor than is its counterpart θ_{ps}^* in the "public signal model." In other words, even if a regime could have survived if all agents believe that a rumor against it is trustworthy, the same regime would collapse when citizens know that the rumor may be uninformative but they are allowed to communicate with one another.

Figure 7 shows such a possibility. In plotting this figure, we choose a value of σ_x^2 lower than that used in Figure 5(a). When z is close to z', the slope of equilibrium θ^* is steeper than θ_{ps}^* (the monotonic decreasing dashed line) at $z = z'$.

To understand this result, note that the sensitivity of θ^* to a change in z around z' is governed by the magnitude of $\partial\hat{x}/\partial\theta$. With more precise private signals, citizens in the "public signal model" would be less responsive to the public information, even though it is fully believed to be informative. Specifically the weight given to private information in the standard linear Bayesian updating formula is β. The value of $\partial\hat{x}_{ps}/\partial\theta$ is equal to $1/\beta$, which increases in σ_x^2; see the dashed line Figure 8. This explains why θ_{ps}^* becomes flatter as σ_x^2 becomes smaller. The same reason can explain why $\partial\hat{x}_m/\partial\theta$ increases; see the dotted line in Figure 8. The widening gap between these two reflects

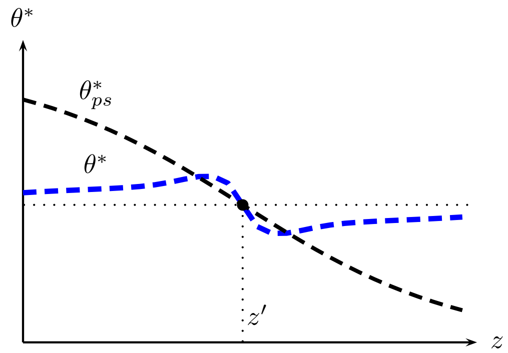


Figure 7. *Comparing the "communication model" to the "public signal model" when private information has reasonably high precision*

the decline in w.

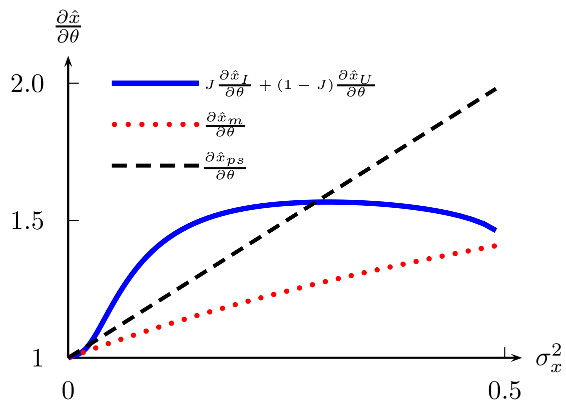


Figure 8. *$\partial\hat{x}/\partial\theta$ evaluated at $x = x'$ and $\theta = \theta'$ in response to σ_x^2, when $z = z'$.*

In the "communication model," an increase in σ_x^2 raises both $\partial\hat{x}_I/\partial\theta$ and $\partial\hat{x}_U/\partial\theta$. However, less weight is attached to the larger value $\partial\hat{x}_I/\partial\theta$ because J falls as σ_x^2 increases. These two opposing effects account for the hump shape of $J\partial\hat{x}_I/\partial\theta + (1 - J)\partial\hat{x}_U/\partial\theta$ shown by the solid curve in Figure 8. For very small or very large values of σ_x^2, this curve is close to $\partial\hat{x}_m/\partial\theta$, and is always above it by equation (23) in the previous subsection. For intermediate values of σ_x^2, the effect of the changing weight J can be so strong that the solid curve rises above $\partial\hat{x}_{ps}/\partial\theta$. That is,

$$J(\theta', z')\frac{\partial\hat{x}_I}{\partial\theta} + (1 - J(\theta', z'))\frac{\partial\hat{x}_U}{\partial\theta} > \frac{\partial\hat{x}_{ps}}{\partial\theta} = \frac{1}{\beta}.$$

Using the same logic as in Proposition 3, this inequality leads to

$$\frac{\mathrm{d}\theta^*(z')}{\mathrm{d}z} < \frac{\mathrm{d}\theta^*_{ps}(z')}{\mathrm{d}z} < 0.$$

This is the situation shown in Figure 7 when σ_x^2 is relatively small.

6. Discussion

6.1. Exchange Private Signals

To this point, we have assumed that communication takes the form of talking about the rumor; that is, citizens communicate their views about the informativeness of the public signal to one another. We show that such communication about publicly observed rumor can make citizens more responsive to the rumor. This kind of communication is qualitatively different from communication about privately observed signals: the former can reinforce the importance of public information while the latter always undermines it. To stress this point, we explore an alternative communication assumption, which lets citizens exchange their private information with one another.

Specifically, in this setting, citizens are randomly paired with one another to exchange their private signals concerning the strength of the regime. To be more general, we allow the communication process to be noisy: each citizen receives her partner's private signal with an additive noise. That is, the message y_i received by the citizen i from her partner j is given by:

$$y_i = x_j + \xi_j,$$

where the noise $\xi_j \sim \mathcal{N}(0, \sigma_\xi^2)$ is normally distributed and independent across j. At the end of communication stage, each citizen possesses an information set which consists of the publicly observed rumor z and two private signals x_i and y_i of different qualities. Given the distributional assumption of the private signals, this information set is equivalent to (z, xy_i), where xy_i is a private signal with higher precision than x_i. That is,

$$\begin{aligned} xy_i &\equiv \frac{\sigma_x^2 + \sigma_\xi^2}{2\sigma_x^2 + \sigma_\xi^2} x_i + \frac{\sigma_x^2}{2\sigma_x^2 + \sigma_\xi^2} y_i, \\ \sigma_{xy}^2 &\equiv \frac{\sigma_x^2 + \sigma_\xi^2}{2\sigma_x^2 + \sigma_\xi^2} \sigma_x^2 < \sigma_x^2. \end{aligned}$$

In other words, this setting is observationally equivalent to the "mute model," with private signals of higher quality. More precise private signals result in two opposing effects for the cut-off types of citizens. First, higher signal quality leads to a higher value of β, which means that citizens will be less responsive to public information (given that it is informative). Second, the weight assigned to the possibility that rumor is informative also changes. For a rumor which is neutral to the regime (i.e., $z = \tilde{z}$), the effect of a change in weight w is zero because the cut-off type citizen does not care whether the rumor is informative or not. Therefore, the first effect dominates.

This explains why exchanging private signals makes the equilibrium threshold regime strength less responsive with respect to the realization of the rumor.

Figure 9 presents the equilibrium threshold regime strengths in the "communication model," the "mute model," and the model with the modified assumption concerning the exchange of private signals (labeled θ^*_{ex}). The same set of parameters is used throughout these examples; the only difference being the communication protocol. The result highlights that the mode of communication matters: when we allow citizens to exchange what they privately know, they rely more on the private signals; when we allow them to exchange views on what they commonly know, they rely more on the public information.

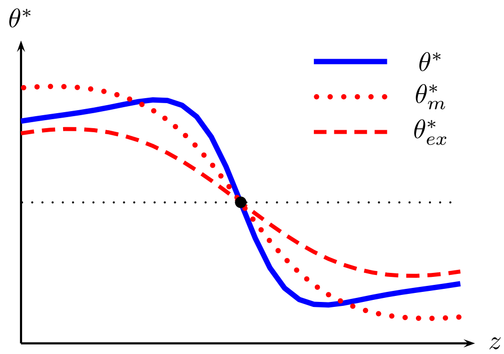


Figure 9. *Equilibrium threshold regime strength is less response to rumors when citizens communicate with one another about their private signals*

6.2. Sentiment

We interpret the mean of the uninformative distribution as sentiment or public perception of what uninformative messages would sound like. Given that the uninformative distribution U has a density that peaks at $z = s$, the weight citizens assign to the possibility that the rumor is informative, $w(z, x_i)$, decreases in the relative distance between z and s. In other words, if s is further away from z, citizens would consider the rumor more likely to be drawn from informative source. If a rumor is close to what the public perceives as untrustworthy information, it will be given less credibility by citizens. Therefore, sentiment critically affects citizens' evaluation of the rumor's informativeness . For example, if citizens have lived with systematic government propaganda for a long time, any claims about its alleged strength would be considered less credible.

Suppose $s = \tilde{z}$. We call this value "neutral sentiment," since equilibrium $\theta^*(z)$ is symmetric around $\tilde{z}$ in this case. If s is exogenously decreased from the neutral value, then for any $z > \tilde{z}$, the equilibrium $\theta^*(z)$ goes down, indicating that the regime is more likely to survive. The key intuition is that when s is low, citizens believe that a rumor

for the regime ($z > \tilde{z}$) is likely to have come from an informative source. As a result, they believe the regime is strong and are therefore less prepared to attack. See Figure 10 for an illustration.

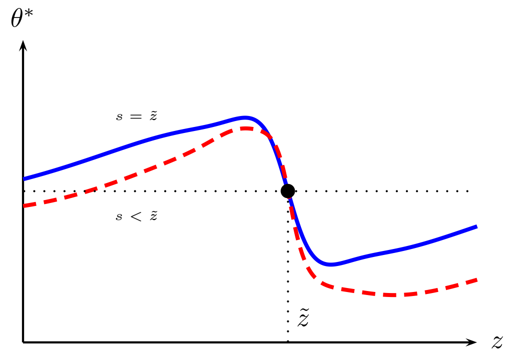


Figure 10. *Lower value of s tends to increase the chance of regime survival*

When s is low (relative to the neutral value) and when the rumor is against the regime ($z < \tilde{z}$), citizens tend to attribute the rumor to an uninformative source. Because they do not believe that the regime is weak, they are again less likely to attack. As is shown in Figure 10, a low value of s tends to increase the chance of regime survival. Thus autocratic regimes that understand the importance of public perception could increase their chances of survival by manipulating it. During the post-election Iranian protest in 2009, the Iranian government deliberately created rumors against itself and then disproved it on national television, which turned out to be effective in discrediting the opposition (Esfandiari 2010).

In reality, sudden shifts in sentiment often take place during revolutions. For instance, the rumor that Mubarak's family had fled Egypt would have sounded ridiculous without the previous successful Tunisian revolution, during which Ben Ali actually escaped. The fact that it was widely reported and believed during the Egyptian unrest, was a sign of a sentiment shift. In this case, the confirmed rumor in Tunisia shifted the sentiment upwards in Egypt and made the Egyptian regime more vulnerable.

6.3. Censorship: The Power of Silence

Our model shows that rumors against a regime could coordinate more citizens to attack, provided the rumors sound plausible. Thus autocratic governments may want to block negative rumors against it and to stop citizens from talking about these rumors. In reality, news censorship and the control of information flow are commonly adopted by many governments concerned about their survival. However since silence itself is a signal about the regime strength, it is not obvious that news censorship necessarily

increases the chances of regime survival. Our model provides a framework that can shed some light on this issue.

Assume that the regime blocks any rumor z if $z < K$. In other words, only rumors that suggest the regime is stronger than K can be heard and discussed by citizens. We assume that citizens are aware of this censorship rule. Therefore it is common knowledge that $z < K$ when citizens do not hear a rumor. In this case, citizens cannot communicate regarding the realization z, given they do not have any knowledge of it.

When z is observable, i.e. $z \geq K$, the equilibrium is exactly the same as in our "communication model." When citizens hear no rumor, they understand that the event $z < K$ has taken place but the authority has censored the news. Taking θ_c as the threshold for regime survival, they calculate the expected payoff of revolt in a Bayesian fashion,

$$\Pr[\theta \leq \theta_c | z < K, x_i] = \frac{\int_{-\infty}^{\theta_c} \left[\alpha\Phi\left(\frac{K-t}{\sigma_z}\right) + (1-\alpha)\Phi\left(\frac{K-s}{\sigma_U}\right)\right] \frac{1}{\sigma_x}\phi\left(\frac{t-x_i}{\sigma_x}\right) \mathrm{d}t}{\int_{-\infty}^{\infty} \left[\alpha\Phi\left(\frac{K-t}{\sigma_z}\right) + (1-\alpha)\Phi\left(\frac{K-s}{\sigma_U}\right)\right] \frac{1}{\sigma_x}\phi\left(\frac{t-x_i}{\sigma_x}\right) \mathrm{d}t}.$$

In this case, the cut-off type citizens with private information x_c^* must be indifferent between revolt and not revolt. The mass of attackers in state θ_c^* must be equal to the threshold regime strength. In other words, the equilibrium pair (θ_c^*, x_c^*) solves the attack function and the associated indifference condition for all $z < K$:

$$\Phi\left(\frac{x_c^* - \theta_c^*}{\sigma_x}\right) = \theta_c^*,$$
$$\Pr\left[\theta \leq \theta_c^* | x_c^*, z < K\right] = c.$$

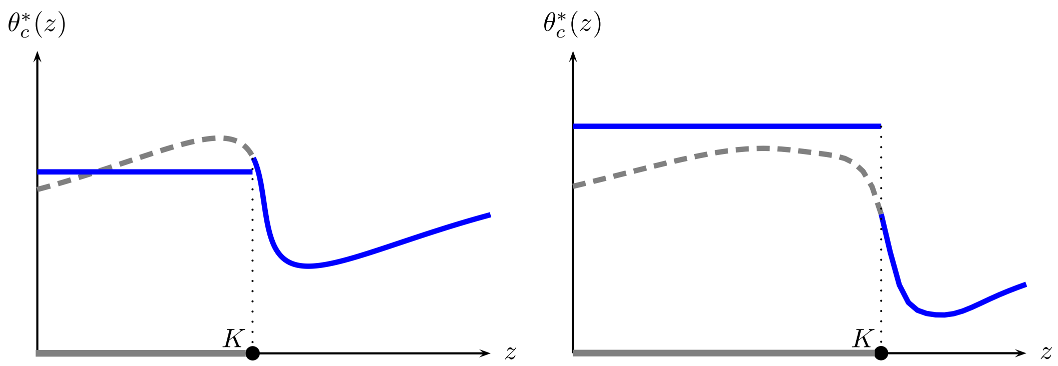
The equilibrium values of θ_c^* depends on the censorship rule K. When K is very small, it corresponds to the case where there is no effective censorship. When K is very high, the regime blocks all the public information, no matter it is for or against it. Therefore, in either case, θ_c^* converges to the value θ_{ms}^* in the "pure noise model."

For intermediate values of the censorship rule, we find that θ_c^* is increasing then decreasing in K. To understand why θ_c^* can be increasing, suppose the censorship rule is raised from an initially low value of K to a slightly higher value K'. This means that rumors with values in $[K, K']$ are now screened out. In a model with trustworthy news only, $z \in [K, K']$ is better news for the regime than $z < K$. So the expected strength of the regime should be higher conditional on $z < K'$ than conditional on $z < K$. In our model, however, rumors with $z \in [K, K']$ can be *worse* news for the regime than those with $z < K$, because the former is deemed plausible while the latter is deemed incredible. This explains why citizens' expectation about the strength of the regime

conditional on hearing no news can be *lower* when the censorship rule is tightened. As the expected payoff from attacking upon hearing no news rises, the equilibrium threshold for regime survival θ_c^* becomes higher.

The effect of censorship on the chances of regime survival depends on the comparison of θ_c^* with $\theta^*(z)$ for $z < K$. There are two scenarios arising from this model. On the one hand, the censorship rule can effectively lower the equilibrium threshold of regime strength (for some values of z) and therefore contribute to the survival of the regime. On the other hand, when private signals are sufficiently imprecise, censorship could backfire.

Figure 11(a) provides an example of the first scenario. The effect of censorship is represented by the difference between the dashed line for the "communication model" and the solid line for the model with censorship. When the rumor is plausible and effectively blocked by the regime, the censorship rule helps the regime achieve better chance of survival by decreasing the equilibrium threshold of regime strength. When the regime blocks a wild rumor, which would be disbelieved by most of citizens if it has reached to all the citizens, censorship does not help the regime. However, the censorship rule could be still worthwhile because, under the assumption that both I and U are normal distributions, extreme rumors are less likely to be generated than are more plausible ones.



(a) Effective censorship

(b) Censorship backfires

Figure 11. *The "censorship model"*

Figure 11(b) illustrates the possibility of the "backfire scenario": when the regime adopts censorship rule K, θ_c^* is higher than $\theta^*(z)$ for all $z < K$. This means that the regime is worse off by adopting censorship. In the "communication model", citizens can observe and discuss the rumor they hear. When the rumor is too extreme (i.e., far below the equilibrium threshold $\theta^*(z)$), communication discourages citizens from attacking and mobilizes fewer attackers than in the "mute model" if θ is near $\theta^*(z)$, as we analyze in subsection 5.2. Shutting out the rumor by censorship also shuts out

this communication effect. Each citizen would then have to make inference from the absence of rumors without the help of what would otherwise be mostly contradictory messages from their peers. In some cases, silence can be more dangerous to the regime than negative but discredited rumors.

7. Conclusion

It is not news that revolutions in history are often intertwined with rumors. It is also not surprising that many rumors are intended to spur individual's participation. However, what strikes us is why those rumors, which often turned out to be false later, could be so effective for mobilization. In this paper, we offer an analysis of this phenomenon in a global game framework. In this model, individuals' skepticism towards rumors arises as a rational response, instead of a behavioral assumption. Moreover, we explicate a novel communication mechanism, where the effect of agents' skepticism can be undone by communication among themselves. When the negative rumor is close to true state of nature, communication on the informativeness of rumors coordinates more citizens than when communication is not allowed. Extensions of our model shed lights on effects of the regime's information manipulation of various forms.

To the best of our knowledge, our model is the first attempt to explicitly investigate the role of rumors in a regime change game. In this paper, we choose not to address how rumors spread across individuals during revolutions. We leave this interesting topic to future work. Moreover, this paper provides a framework using mixture distributions to study the transmission of unreliable information between agents engaged in collective actions. Such a framework may be useful in other applications. Our theory is interpreted in the context of political revolution, but it can also be extended to model rumors in bank runs, financial crises or currency attacks.

Appendix

Proof of Proposition 1. In the "pure noise model," the indifference condition for the cut-off type x^*_{ms} satisfies

$$\Phi\left(\frac{\theta^*_{ms} - x^*_{ms}}{\sigma_x}\right) = c,$$

while attack equation for the threshold regime strength θ^*_{ms} satisfies

$$\Phi\left(\frac{x^*_{ms} - \theta^*_{ms}}{\sigma_x}\right) = \theta^*_{ms}.$$

Solving these two equations gives the equilibrium values of θ^*_{ms} and x^*_{ms} as stated in the proposition.

In the "public signal model," let $X^*_{ps} = \beta x^*_{ps} + (1-\beta)z$ denote the posterior mean of θ for the cut-off type x^*_{ps}. The posterior variance is $\beta\sigma_x^2$. The indifference condition and the attack equation can be written, respectively, as:

$$\Phi\left(\frac{\theta^*_{ps} - X^*_{ps}}{\sqrt{\beta}\sigma_x}\right) = c, \tag{25}$$

$$\Phi\left(\frac{x^*_{ps} - \theta^*_{ps}}{\sigma_x}\right) = \theta^*_{ps}. \tag{26}$$

Eliminate x^*_{ps} from these two equations to get

$$\theta^*_{ps} - \frac{\beta\sigma_x}{(1-\beta)}\Phi^{-1}(\theta^*_{ps}) = \frac{\sqrt{\beta}\sigma_x}{(1-\beta)}\Phi^{-1}(c) + z. \tag{27}$$

Condition (1) implies that the left-hand-side of the above is decreasing in θ^*_{ps}. Since the right-hand-side is increasing in z, we have $\mathrm{d}\theta^*_{ps}/\mathrm{d}z < 0$. From the indifference condition and the definition of X^*_{ps}, we have

$$\frac{\mathrm{d}\theta^*_{ps}}{\mathrm{d}z} - \beta\frac{\mathrm{d}x^*_{ps}}{\mathrm{d}z} - (1-\beta) = 0.$$

Hence $\mathrm{d}x^*_{ps}/\mathrm{d}z < 0$. This establishes part (1).

Taking limit of equation (27) gives $\lim_{z\to\infty}\theta^*_{ps}(z) = 0$. The attack equation then implies $\lim_{z\to\infty} x^*_{ps}(z) = -\infty$. This establishes part (2a). Part (2b) is obtained analogously.

Finally, since $\lim_{z\to\infty}\theta^*_{ps}(z) < \theta^*_{ms} < \lim_{z\to-\infty}\theta^*_{ps}(z)$, and since $\theta^*_{ps}(z)$ is strictly decreasing, there exists a unique $\tilde{z}$ such that $\theta^*_{ps}(\tilde{z}) = \theta^*_{ms}$. From the attack equations in the "pure noise model" and in the "public signal model," $\theta^*_{ps} = \theta^*_{ms}$ implies $x^*_{ps} = x^*_{ms}$.

Hence, $x^*_{ps}(\tilde{z}) = x^*_{ms}$. This establishes part (3). ■

Proof of Proposition 2. *Part (1).* Since $(\theta^*_{ms}, x^*_{ms})$ satisfies the indifference condition of the "pure noise model," we have

$$\Phi\left(\frac{\theta^*_{ms} - x^*_{ms}}{\sigma_x}\right) = c.$$

By part (3) of Proposition 1, $(\theta^*_{ms}, x^*_{ms})$ also satisfies the indifference condition of the "public signal model" when $z = \tilde{z}$. Therefore,

$$\Phi\left(\frac{\theta^*_{ms} - (\beta x^*_{ms} + (1-\beta)\tilde{z})}{\sqrt{\beta}\sigma_x}\right) = c.$$

The posterior belief $P(\theta^*_{ms}|\tilde{z}, x^*_{ms})$ in the "mute model" is just the weighted average of the left-hand-side of the two equations above. Hence $P(\theta^*_{ms}|\tilde{z}, x^*_{ms}) = c$. Furthermore, $(\theta^*_{ms}, x^*_{ms})$ satisfies the attack equation, $\Phi\left(\sigma_x^{-1}(x^*_{ms} - \theta^*_{ms})\right) = \theta^*_{ms}$. Therefore, it is an equilibrium in the "mute model" for $z = \tilde{z}$.

Part (2). We first show that if $z < \tilde{z}$, then $\theta^*_m(z) < \theta^*_{ps}(z)$. From the attack equation (15), we can write $\Phi\left(\sigma_x^{-1}(\theta^*_m - x^*_m)\right) = 1 - \theta^*_m$. Therefore, the indifference condition (14) of the "mute model" can be written as:

$$c = w(z, x^*_m)\Phi\left(\frac{\theta^*_m - (\beta x^*_m + (1-\beta)z)}{\sqrt{\beta}\sigma_x}\right) + (1 - w(z, x^*_m))(1 - \theta^*_m). \tag{28}$$

From the indifference condition of the "public signal model" and from the fact that $1 - \theta^*_{ms} = c$, we also have

$$c = w(z, x^*_m)\Phi\left(\frac{\theta^*_{ps} - (\beta x^*_{ps} + (1-\beta)z)}{\sqrt{\beta}\sigma_x}\right) + (1 - w(z, x^*_m))(1 - \theta^*_{ms}).$$

These two equations, together with the fact that $\theta^*_{ps} > \theta^*_{ms}$ when $z < \tilde{z}$, imply

$$\begin{aligned} & w(z, x^*_m)\Phi\left(\frac{q(\theta^*_m) - (1-\beta)z}{\sqrt{\beta}\sigma_x}\right) + (1 - w(z, x^*_m))(1 - \theta^*_m) \\ > \; & w(z, x^*_m)\Phi\left(\frac{q(\theta^*_{ps}) - (1-\beta)z}{\sqrt{\beta}\sigma_x}\right) + (1 - w(z, x^*_m))(1 - \theta^*_{ps}), \end{aligned}$$

where

$$q(\theta^*) \equiv \theta^* - \beta x^* = \theta^* - \beta(\theta^* + \sigma_x \Phi^{-1}(\theta^*)),$$

because the equilibrium x^* in both the "mute model" and the "public signal model"

must satisfy the attack equation. To show $\theta_m^* < \theta_{ps}^*$ from the above inequality, it suffices to show that $\mathrm{d}q(\theta^*)/\mathrm{d}\theta^* \leq 0$. We have

$$\frac{\mathrm{d}q(\theta^*)}{\mathrm{d}\theta^*} = 1 - \beta - \frac{\beta\sigma_x}{\phi\left(\Phi^{-1}(\theta^*)\right)} \leq 1 - \beta - \beta\sigma_x\sqrt{2\pi},$$

which is negative by assumption (1). Hence $\theta_m^* < \theta_{ps}^*$ when $z < \tilde{z}$.

To establish part (2a), we also need to show that $\theta_m^*(z) > \theta_{ms}^*$ if $z < \tilde{z}$. Suppose this is not true. Then $1 - \theta_m^* \geq 1 - \theta_{ms}^* = c$. Therefore equation (28) implies

$$\Phi\left(\frac{\theta_m^* - (\beta x_m^* + (1-\beta)z)}{\sqrt{\beta}\sigma_x}\right) \leq c = \Phi\left(\frac{\theta_{ms}^* - (\beta x_{ms}^* + (1-\beta)\tilde{z})}{\sqrt{\beta}\sigma_x}\right),$$

where the equality follows because $(\theta_{ms}^*, x_{ms}^*)$ satisfies the indifference condition of the "public signal model" at $z = \tilde{z}$. From the above inequality, we obtain

$$q(\theta_m^*) - q(\theta_{ms}^*) \leq (1-\beta)(z - \tilde{z}) < 0.$$

Since $\mathrm{d}q(\theta^*)/\mathrm{d}\theta^* < 0$, this inequality implies $\theta_m^* > \theta_{ms}^*$, a contradiction. Thus, when $z < \tilde{z}$, we must have $\theta_m^*(z) \in (\theta_{ms}^*, \theta_{ps}^*(z))$. The proof of part (2b) is symmetric.

Part (3). Combine equations (14) and (15) to get

$$\theta_m^* = (1-c) + \left[\Phi\left(\frac{\theta_m^* - [\beta x_m^* + (1-\beta)z]}{\sqrt{\beta}\sigma_x}\right) - c\right] \cdot \frac{w(z, x_m^*)}{1 - w(z, x_m^*)}.$$

As z goes to infinity, $x_m^*(z)$ must remain finite, otherwise it would violate equation (15). For any finite x_m^*, $\lim_{z\to\infty} w(z, x_m^*) = 0$. Therefore,

$$\lim_{z\to\infty} \theta_m^*(z) = 1 - c = \theta_{ms}^*.$$

A similar argument establishes that $\lim_{z\to-\infty} \theta_m^*(z) = \theta_{ms}^*$.

Part (4). We show that (a) $\theta_m^*(z)$ is increasing then decreasing for $z \in (-\infty, \tilde{z}]$; and (b) $\theta_m^*(z)$ is decreasing then increasing for $z \in [\tilde{z}, \infty)$.

Fix a $z_0 \in (-\infty, \tilde{z}]$. Define

$$f(z) \equiv P(\theta_m^*(z_0)|z, x_m^*(z_0))$$

$z \in (-\infty, \tilde{z}]$. We show that f is single-peaked in z. It suffices to verify that $\mathrm{d}f(z)/\mathrm{d}z = 0$ implies $\mathrm{d}^2 f(z)/\mathrm{d}z^2 < 0$. To simplify the notation, let Φ_I represent the standard normal distribution evaluated at the point $(\theta_m^*(z_0) - \beta x_m^*(z_0) - (1-\beta)z)/(\sqrt{\beta}\sigma_x)$, and

let Φ_U represent the standard normal distribution evaluated at the point $(\theta_m^*(z_0) - x_m^*(z_0))/\sigma_x$. Note that Φ_I depends on z but Φ_U does not. Define ϕ_I and ϕ_U analogously. We have

$$\begin{aligned}\frac{\mathrm{d}f(z)}{\mathrm{d}z} &= (\Phi_I - \Phi_U)\frac{\partial w}{\partial z} - w\frac{\partial \Phi_I}{\partial z} \\ &= \left[(1-w)\left(\frac{z-s}{\sigma_U^2} + \frac{x_m^*(z_0)-z}{\sigma_I^2}\right)\left(1-\frac{\Phi_U}{\Phi_I}\right) - \frac{1-\beta}{\sqrt{\beta}\sigma_x}\frac{\phi_I}{\Phi_I}\right] w\Phi_I.\end{aligned}$$

When $\mathrm{d}f(z)/\mathrm{d}z = 0$, the second derivative is given by:

$$\begin{aligned}\frac{1}{w\Phi_I}\frac{\mathrm{d}^2 f(z)}{\mathrm{d}z^2} &= -w(1-w)\left(\frac{z-s}{\sigma_u^2} + \frac{x_m^*(z_0)-z}{\sigma_I^2}\right)^2\left(1-\frac{\Phi_U}{\Phi_I}\right) \\ &\quad + (1-w)\left(\frac{1}{\sigma_U^2} - \frac{1}{\sigma_I^2}\right)\left(1-\frac{\Phi_U}{\Phi_I}\right) \\ &\quad + (1-w)\left(\frac{z-s}{\sigma_U^2} + \frac{x_m^*(z_0)-z}{\sigma_I^2}\right)\left(\frac{\mathrm{d}\left(1-\Phi_U/\Phi_I\right)}{\mathrm{d}z}\right) \\ &\quad - \left(\frac{1-\beta}{\sigma_x\sqrt{\beta}}\right)\frac{\mathrm{d}\left(\phi_I/\Phi_I\right)}{\mathrm{d}z}.\end{aligned}$$

(i) The first term is negative because $\Phi_I > \Phi_U$ for $z < \tilde{z}$. To see this, suppose the contrary is true. Then $\Phi_I \leq \Phi_U$ implies

$$\theta_m^*(z_0) + \sqrt{\beta}x_m^*(z_0) \leq (1+\sqrt{\beta})z.$$

But for $z_0 \leq \tilde{z}$ the left-hand-side is greater than $\theta_{ms}^* + \sqrt{\beta}x_{ms}^*$, which is equal to $(1+\sqrt{\beta})\tilde{z}$, a contradiction. (ii) The second term is negative because $\sigma_U^2 > \sigma_I^2$ from assumption (2). (iii) The third term is negative because $(z-s)/\sigma_U^2 + (x_m^*(z_0)-z)/\sigma_I^2 > 0$ whenever $\mathrm{d}f(z)/\mathrm{d}z = 0$ and because $1-\Phi_U/\Phi_I$ is decreasing in z. (iv) The fourth term is negative because the function ϕ_I/Φ_I is increasing in z.

The single-peakedness of $f(z)$ for $z \in (-\infty, \tilde{z}]$ implies that in this range there can be at most one $z_1 \neq z_0$ such that $\theta_m^*(z_1) = \theta_m^*(z_0)$. Suppose otherwise. Let $z_1 \neq z_2 \neq z_0$ be such that $\theta_m^*(z_1) = \theta_m^*(z_2) = \theta_m^*(z_0)$. By the attack equation (15), this implies $x_m^*(z_1) = x_m^*(z_2) = x_m^*(z_0)$. Since the pair $(\theta_m^*(z_0), x_m^*(z_0))$ satisfy the equilibrium conditions for $z \in \{z_1, z_2, z_0\}$, the equation $f(z) = c$ has at least three solutions, which contradicts the single-peakedness of f.

Parts (1), (2a), and (3) of the proposition, coupled with the fact that there is at most one $z_1 \neq z_0$ such that $\theta_m^*(z_1) = \theta_m^*(z_0)$, together establish that $\theta_m^*(z)$ is increasing then decreasing for $z \in (-\infty, \tilde{z}]$.

For the case $z \in [\tilde{z}, \infty)$, write:

$$\frac{\mathrm{d}f(z)}{\mathrm{d}z} = \left[(1-w)\left(\frac{z-s}{\sigma_U^2} + \frac{x_m^*(z_0)-z}{\sigma_I^2}\right)\left(\frac{1-\Phi_U}{1-\Phi_I} - 1\right) - \frac{1-\beta}{\sqrt{\beta}\sigma_x}\frac{\phi_I}{1-\Phi_I}\right] w(1-\Phi_I).$$

We can show that the bracketed term is increasing when it is equal to zero, because $\Phi_I < \Phi_U$ when $z > \tilde{z}$ and because $\phi_I/(1-\Phi_I)$ is decreasing in z. Hence $f(z)$ must be decreasing then increasing in z in this range, which, together with parts (1), (2b), and (3) of the proposition, imply that $\theta_m^*(z)$ is decreasing then increasing for $z \in [\tilde{z}, \infty)$. ■

Proof of Proposition 3. *Part (1).* We only prove the limiting properties of $x_U^*(z)$ and $\theta^*(z)$ here. The limiting properties of $x_I^*(z)$ is characterized in Lemma 1 in the Technical Appendix.

Consider the following equation system:

$$\lim_{z\to\infty} \int_{-\infty}^{\theta} p(t|z,x)\,\mathrm{d}t = c,$$
$$\Phi\left(\frac{x-\theta}{\sigma_x}\right) = \theta.$$

We note that, by Proposition 2, $(\theta_{ms}^*, x_{ms}^*)$ is the unique solution to this system.

Consider the "communication model" now. Suppose the limit values of both $x_U^*(z)$ and $\theta^*(z)$ are finite. The indifference condition (10) in "communication model" implies

$$\lim_{z\to\infty} \int_{-\infty}^{\theta^*} \frac{1-J(t,z)}{\Pr[y_i = 0|z, x_U^*]} p(t|z, x_U^*)\,\mathrm{d}t = c.$$

By Lemma 1 (Claim 1) in the Technical Appendix, both $\underline{x}(z)$ and $\overline{x}(z)$ go to infinity as z goes to infinity. Therefore, for any $t \leq \theta^*$, the probability that x_j does not belong to $[\underline{x}(z), \overline{x}(z)]$ goes to one. We thus have

$$\lim_{z\to\infty} \frac{1-J(t,z)}{\Pr[y_i = 0|z, x_U^*]} = 1.$$

Therefore indifference condition (10) for the "communication model" becomes,

$$\lim_{z\to\infty} \int_{-\infty}^{\theta^*} p(t|z, x_U^*)\,\mathrm{d}t = c.$$

When z goes to infinity, $J(\theta^*, z)$ goes to zero. Therefore the attack equation (11) for

the "communication model" becomes,

$$\lim_{z\to\infty} J(\theta^*,z)\Phi\left(\frac{x_I^*-\theta^*}{\sigma_x}\right)+(1-J(\theta^*,z))\Phi\left(\frac{x_U^*-\theta^*}{\sigma_x}\right)=\Phi\left(\frac{x_U^*-\theta^*}{\sigma_x}\right)=\theta^*.$$

Given (θ^*, x_U^*) solves the same equation system as that in the "pure noise model," we conclude that $\lim_{z\to\infty} x_U^*(z) = x_{ms}^*$ and $\lim_{z\to\infty} \theta^*(z) = \theta_{ms}^*$. The proof of the case for the limit as z goes to minus infinity is analogous.

Part (2). From part (1), $x_I^*(z) > x_U^*(z)$ for z sufficiently small and $x_I^*(z) < x_U^*(z)$ for z sufficiently large. Both $x_I^*(z)$ and $x_U^*(z)$ are continuous. Therefore there exists a z' such that $x_I^*(z') = x_U^*(z')$.

Let $\theta^*(z') = \theta'$ and $x_I^*(z') = x_U^*(z') = x'$. We proceed to establish that (θ', x') solves the "mute model" as well. To see this, we first note that $x_I^* = x_U^*$ implies that the mass of the swing population is zero. Hence, the attack equation (11) of the "communication model" reduces to the attack equation (15) of the "mute model." Next, note that for any value of z', x' and θ', we have

$$P(\theta'|z',x') = \Pr[y_i = 1|z',x']P(\theta'|z',x',1) + \Pr[y_i = 0|z',x']P(\theta'|z',x',0).$$

Therefore, if z', x', and θ' satisfy the indifference conditions $P(\theta'|z',x',1) = c$ and $P(\theta'|z',x',0) = c$ in the "communication model," then they must satisfy the indifference condition $P(\theta'|z',x') = c$ in the "mute model" as well.

Part (3). Let $\theta_m^*(\tilde{z}) = \tilde{\theta}$ and $x_m^*(\tilde{z}) = \tilde{x}$. We need to show that $(\tilde{\theta}, \tilde{x}, \tilde{x})$ solves the "communication model" at $z = \tilde{z}$. To this end, it suffices to show that $P(\tilde{\theta}|\tilde{z},\tilde{x},1) = c$. Given that $(\tilde{\theta}, \tilde{x})$ solves the "mute model" at $z = \tilde{z}$, we have $P(\tilde{\theta}|\tilde{z},\tilde{x}) = c$. These two conditions would imply that $P(\tilde{\theta}|\tilde{z},\tilde{x},0) = c$. Hence the indifference condition (10) for the "communication model" is satisfied. Given that $x_I^*(\tilde{z}) = x_U^*(\tilde{z}) = \tilde{x}$, the attack equation (11) holds as well.

The indifference condition $P(\tilde{\theta}|\tilde{z},\tilde{x},1) = c$ can be rewritten as follows,

$$w(T_1 - cT_2) + (1-w)(T_3 - cT_4) = 0,$$

where $T_1 = \Pr[\theta \le \hat{\theta}, y_i = 1|x_I^*, z, z \sim I]$, $T_2 = \Pr[y_i = 1|x_I^*, z, z \sim I]$, $T_3 = \Pr[\theta \le \hat{\theta}, y_i = 1|x_I^*]$, and $T_4 = \Pr[y_i = 1|x_I^*]$. Specifically,

$$T_1 = \int_{-\infty}^{\tilde{\theta}} J(t,\tilde{z})\frac{1}{\sqrt{\beta}\sigma_x}\phi\left(\frac{t-\tilde{x}-(1-\beta)\tilde{z}}{\sqrt{\beta}\sigma_x}\right)\mathrm{d}t.$$

For $c = 0.5$, $\tilde{x} = \tilde{\theta} = \tilde{z}$. Therefore, using the substitution $t' = t - \tilde{\theta}$, we obtain

$$T_1 = \int_{-\infty}^{0} J(t' + \tilde{z}, \tilde{z}) \frac{1}{\sqrt{\beta}\sigma_x} \phi\left(\frac{t'}{\sqrt{\beta}\sigma_x}\right) \mathrm{d}t'.$$

Both $J(t' + \tilde{z}, \tilde{z})$ and $\phi(t'/(\sqrt{\beta}\sigma_x))$ are symmetric about the point $t' = 0$. Therefore,

$$T_1 = 0.5 \int_{-\infty}^{\infty} J(t' + \tilde{z}, \tilde{z}) \frac{1}{\sqrt{\beta}\sigma_x} \phi\left(\frac{t'}{\sqrt{\beta}\sigma_x}\right) \mathrm{d}t' = 0.5 T_2.$$

This shows that $T_1 - cT_2 = 0$ when $c = 0.5$. Similarly,

$$T_3 = \int_{-\infty}^{\tilde{\theta}} J(t, \tilde{z}) \frac{1}{\sigma_x} \phi\left(\frac{t - \tilde{x}}{\sigma_x}\right) \mathrm{d}t = 0.5 \int_{-\infty}^{\infty} J(t' + \tilde{z}, \tilde{z}) \frac{1}{\sigma_x} \phi\left(\frac{t'}{\sigma_x}\right) \mathrm{d}t' = 0.5 T_4.$$

This shows that $T_3 - cT_4 = 0$ when $c = 0.5$. Thus, we have $P(\tilde{\theta}|\tilde{z}, \tilde{x}, 1) = c$. ■

Proof of Proposition 4. Write the relevant indifference conditions in the following form:

$$\begin{aligned}
\tau_m(\theta, z, x) &\equiv P(\theta|z, x) - c = 0, \\
\tau_I(\theta, z, x) &\equiv \int_{-\infty}^{\theta} J(t, z) p(t|z, x) \,\mathrm{d}t - c \Pr[y_i = 1|z, x] = 0, \\
\tau_U(\theta, z, x) &\equiv \int_{-\infty}^{\theta} (1 - J(t, z)) p(t|z, x) \,\mathrm{d}t - c(1 - \Pr[y_i = 1|z, x]) = 0.
\end{aligned}$$

From the proof of Lemma 3 (Claim 2) in the Technical Appendix, we have

$$\begin{aligned}
\frac{\partial \tau_I}{\partial x} &= J(\theta', z') \frac{\partial \tau_m}{\partial x} - D, \\
\frac{\partial \tau_U}{\partial x} &= (1 - J(\theta', z')) \frac{\partial \tau_m}{\partial x} + D,
\end{aligned}$$

where $D < 0$.

Let p denote $p(\theta'|z', x')$ and let P_x denote $\partial P(\theta'|z', x')/\partial x$. With the implicit function theorem, we obtain:

$$\begin{aligned}
\frac{\partial \hat{x}_I}{\partial \theta} &= \frac{-Jp}{JP_x - D} > 0, \\
\frac{\partial \hat{x}_U}{\partial \theta} &= \frac{-(1 - J)p}{(1 - J)P_x + D} > 0, \\
\frac{\partial \hat{x}_m}{\partial \theta} &= \frac{-p}{P_x} > 0.
\end{aligned}$$

A direct calculation gives:

$$J\frac{\partial \hat{x}_I}{\partial \theta} + (1-J)\frac{\partial \hat{x}_U}{\partial \theta} - \frac{\partial \hat{x}_m}{\partial \theta} = \frac{-pD^2}{P_x(JP_x - D)((1-J)P_x + D)} > 0.$$

This establishes the first inequality of (23). Moreover, since

$$P_x = -p + (1-\beta)w\phi\left(\frac{\theta' - X'}{\sqrt{\beta}\sigma_x}\right) < 0,$$

we have $\partial \hat{x}_m / \partial \theta > 1$. This establishes the second inequality of (23).

The inequality (24) follows from (23) and from Lemma 3, which shows that

$$\frac{\partial \hat{x}_I}{\partial z} + \frac{\partial \hat{x}_I}{\partial \theta} = \frac{\partial \hat{x}_m}{\partial z} + \frac{\partial \hat{x}_m}{\partial \theta} = \frac{\partial \hat{x}_U}{\partial z} + \frac{\partial \hat{x}_U}{\partial \theta} = 1.$$

Finally, inequalities (24) and (23), together with the comparison of the decomposition equation (17) of the "mute model" and the corresponding decomposition equation (22) of the "communication model," establish the proposition. ■

References

Allport, G.W. and L.J. Postman (1947). *The Psychology of Rumor*. New York,: H. Holt and Company.

Angeletos, G.M. and I. Werning (2006). Crises and prices: Information aggregation, multiplicity, and volatility. *American Economic Review 96*(5), pp. 1720–1736.

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

Bernhardt, D. and M. Shadmehr (2010). Collective Action with Uncertain Payoffs: Coordination, Public Signals and Punishment Dilemmas. *SSRN eLibrary*.

Bilefsky, D. (2009). Celebrating revolution with roots in a rumor. *The New York Times*, November 18, p. 16.

Boix, C. and M. Svolik (2010). The Foundations of Limited Authoritarian Government: Institutions and Power-Sharing in Dictatorships. *SSRN eLibrary*.

Bommel, J.V. (2003). Rumors. *Journal of Finance 58*(4), pp. 1499–1519.

Buchner, H.T. (1965). A theory of rumor transmission. *Public Opinion Quarterly 29*(1), 54–70.

Bueno de Mesquita, E. (2010). Regime change and revolutionary entrepreneurs. *American Political Science Review 104*(03), 446–466.

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

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

Edmond, C. (2011). Information manipulation, coordination, and regime change. Working Paper 17395, National Bureau of Economic Research.

Elias, N. and J.L. Scotson (1994). *The Established and The Outsiders: A Sociology Enquiry into Community Problems*, 2nd ed. SAGE Publications.

Esfandiari, G. (2010). The Twitter devolution. *Foreign Policy*, June 7.

Gambetta, D. (1994). Godfather's gossip. *European Journal of Sociology 35*(02), 199–223.

Gentzkow, M. and J.M. Shapiro (2006). Media bias and reputation. *Journal of Political Economy 114*(2), 280–316.

Goldstone, J.A. (1994). *Revolutions: Theoretical, Comparative, and Historical Studies*, 2d. ed. Fort Worth: Harcourt Brace College Publishers.

Grunden, W.E., M. Walker, and M. Yamazaki (2005). Wartime nuclear weapons research in Germany and Japan. *Osiris 20*, pp. 107–130.

Knapp, R.H. (1944). A psychology of rumor. *The Public Opinion Quarterly 8*(1), pp. 22–37.

Ley, R. (1997). Köhler and espionage on the Island of Tenerife: A rejoinder to Teuber. *American Journal of Psychology 110*(2), pp. 277–284.

Milgrom, P. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics 12*(2), pp. 380–391.

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

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

Nkpa, N.K.U. (1977). Rumors of mass poisoning in Biafra. *Public Opinion Quarterly 41*(3), 332–346.

Peterson, W.A. and N.P. Gist (1951). Rumor and public opinion. *American Journal of Sociology 57*(2), pp. 159–167.

Skocpol, T. (1979). *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. New York: Cambridge University Press.

Suen, W. (2010). Mutual admiration clubs. *Economic Inquiry 48*(1), 123–132.

Turner, M.E., A.R. Pratkanis, P. Probasco, and C. Leve (1992). Threat, cohesion, and group effectiveness: Testing a social identity maintenance perspective on groupthink. *Journal of Personality and Social Psychology 63*(5), 781–796.

World-Tribune.com (Jan 28, 2011). Opposition: Mubarak's family has left Egypt.

Zhang, N. (2009). *Rumors and mobilization in china's contentions politics*. Ph.D. dissertation, Chinese University of Hong Kong.

Technical Appendix

(Not intended for publication)

Lemma 1. In the "communication model," $\lim_{z\to-\infty} x_I^*(z) = +\infty$ and $\lim_{z\to\infty} x_I^*(z) = -\infty$.

Proof. We only prove the first part of the proposition; the proof of the second part is similar.

Claim 1. (a) $\lim_{z\to-\infty} \underline{x}(z)/z = 1 + \sigma_I/\sigma_U$; and (b) $\lim_{z\to-\infty} \overline{x}(z)/z = 1 - \sigma_I/\sigma_U$.

Recall that $\overline{x}(z)$ is the larger solution to $w(z,x) = \delta$, where w is given by equation (4). Solving this equation gives

$$\overline{x}(z) = z + \sqrt{\frac{\sigma_I^2}{\sigma_U^2}(z-s)^2 - 2\sigma_I^2 \log \frac{\sigma_I^2\delta(1-\alpha)}{\sigma_U^2(1-\delta)\alpha}} \equiv z + \kappa(z).$$

Since $\lim_{z\to\infty} \kappa(z)/z = \sigma_I/\sigma_U$, this establishes part (a). Part (b) also follows because $\underline{x}(z) = z - \kappa(z)$.

Claim 2. For any threshold value of $\hat{\theta}$,

$$P(\hat{\theta}|z, x_I^*, y_i = 1) \geq \min\left\{P(\hat{\theta}|z, x_I^*, y_i = 1, z \sim I), P(\hat{\theta}|x_I^*, y_i = 1)\right\}.$$

Let $T_1 = \Pr[\theta \leq \hat{\theta}, y_i = 1|x_I^*, z, z \sim I]$ and $T_2 = \Pr[y_i = 1|x_I^*, z, z \sim I]$. Similarly, let $T_3 = \Pr[\theta \leq \hat{\theta}, y_i = 1|x_I^*]$ and $T_4 = \Pr[y_i = 1|x_I^*]$. Using the weight function w, we can write,

$$P(\hat{\theta}|z, x_I^*, 1) = \frac{wT_1 + (1-w)T_3}{wT_2 + (1-w)T_4}.$$

Note that $T_1/T_2 \geq T_3/T_4$ implies $P(\hat{\theta}|z, x_I^*, 1) \geq T_3/T_4$, and $T_1/T_2 \leq T_3/T_4$ implies $P(\hat{\theta}|z, x_I^*, 1) \geq T_1/T_2$. By the definition of conditional probability, $T_1/T_2 = P(\hat{\theta}|z, x_I^*, y_i = 1, z \sim I)$ and $T_3/T_4 = P(\hat{\theta}|x_I^*, y_i = 1)$. Therefore, the claim follows.

Claim 3. For any finite x_I^* and finite $\hat{\theta}$, $\lim_{z\to-\infty} P(\hat{\theta}|z, x_I^*, y_i = 1, z \sim I) = 1$.

Consider the complementary probability,

$$\Pr[\theta > \hat{\theta}|z, x_I^*, y_i = 1, z \sim I] = \frac{\int_{\hat{\theta}}^{\infty} J(t,z)\frac{1}{\sqrt{\beta}\sigma_x}\phi\left(\frac{t-z-\beta(x_I^*-z)}{\sqrt{\beta}\sigma_x}\right)\mathrm{d}t}{\Phi\left(\frac{\overline{x}(z)-z-\beta(x_I^*-z)}{\sqrt{1+\beta}\sigma_x}\right) - \Phi\left(\frac{\underline{x}(z)-z-\beta(x_I^*-z)}{\sqrt{1+\beta}\sigma_x}\right)}.$$

Since $J(t,z)$ is decreasing in t for $t > z$, we have

$$\lim_{z\to-\infty} \Pr[\theta > \hat{\theta}|z, x_I^*, y_i = 1, z \sim I] \leq \lim_{z\to-\infty} \frac{J(\hat{\theta},z)\left(1-\Phi\left(\frac{\hat{\theta}-z-\beta(x_I^*-z)}{\sqrt{\beta}\sigma_x}\right)\right)}{\Phi\left(\frac{\overline{x}(z)-z-\beta(x_I^*-z)}{\sqrt{1+\beta}\sigma_x}\right)-\Phi\left(\frac{\underline{x}(z)-z-\beta(x_I^*-z)}{\sqrt{1+\beta}\sigma_x}\right)}.$$

Note that $\lim_{z\to-\infty}(\overline{x}(z)-z-\beta(x-z))/z = \beta - \sigma_I/\sigma_U$. There are two cases to consider. If $\beta - \sigma_I/\sigma_U \leq 0$, then the denominator of the above term does not vanish as z goes to minus infinity, while the numerator goes to zero. So the limit of the ratio is 0. If $\beta - \sigma_I/\sigma_U > 0$ then both denominator and numerator vanishes. However, $J(\hat{\theta},z)$ goes to 0 at the rate at which $(\overline{x}(z)-\hat{\theta})/\sigma_x$ goes to minus infinity, which is equal to $(1-\sigma_I/\sigma_U)/\sigma_x$. The denominator goes to 0 at the rate at which $(\overline{x}(z)-z-\beta(x-z))/(\sqrt{1+\beta}\sigma_x)$ goes to minus infinity, which is

$$\frac{\beta-\sigma_I/\sigma_U}{\sqrt{1+\beta}\sigma_x} < \frac{1-\sigma_I/\sigma_U}{\sigma_x}.$$

Hence, in both cases,

$$\lim_{z\to-\infty} \frac{J(\hat{\theta},z)}{\Phi\left(\frac{\overline{x}(z)-z-\beta(x_I^*-z)}{\sqrt{1+\beta}\sigma_x}\right)-\Phi\left(\frac{\underline{x}(z)-z-\beta(x_I^*-z)}{\sqrt{1+\beta}\sigma_x}\right)} = 0.$$

This implies that $\lim_{z\to-\infty} P(\hat{\theta}|z, x_I^*, y_i = 1, z \sim I) = 1$.

Claim 4. For any finite x_I^* and finite $\hat{\theta}$, $\lim_{z\to-\infty} P(\hat{\theta}|x_I^*, y_i = 1) = 1$.

Consider limit of the complementary probability,

$$\begin{aligned}\lim_{z\to-\infty} \Pr[\theta > \hat{\theta}|x_I^*, y_i = 1] &= \lim_{z\to\infty} \frac{\int_{\hat{\theta}}^{\infty} J(t,z)\frac{1}{\sigma_x}\phi\left(\frac{t-x_I^*}{\sigma_x}\right)\,\mathrm{d}t}{\Phi\left(\frac{\overline{x}(z)-x_I^*}{\sqrt{2}\sigma_x}\right)-\Phi\left(\frac{\underline{x}(z)-x_I^*}{\sqrt{2}\sigma_x}\right)} \\ &\leq \lim_{z\to-\infty} \frac{J(\hat{\theta},z)\left(1-\Phi\left(\frac{\hat{\theta}-x_I^*}{\sigma_x}\right)\right)}{\Phi\left(\frac{\overline{x}(z)-x_I^*}{\sqrt{2}\sigma_x}\right)-\Phi\left(\frac{\underline{x}(z)-x_I^*}{\sqrt{2}\sigma_x}\right)}.\end{aligned}$$

Note that $J(\hat{\theta},z)$ goes to 0 at the rate $(1-\sigma_I/\sigma_U)/\sigma$, while the denominator goes to 0 at the slower rate of $(1-\sigma_I/\sigma_U)/(\sqrt{2}\sigma_x)$. Hence the limit of the ratio is 0, and we have $\lim_{z\to-\infty} P(\hat{\theta}|x_I^*, y_i = 1) = 1$.

Claim 5. $\lim_{z\to-\infty} x_I^*(z)$ is not finite.

If $x_I^*(z)$ is finite, by Claims 2 to 4 above, we have

$$\lim_{z \to -\infty} P(\hat{\theta}|z, x_I^*, y_i = 1) = 1 > c,$$

for any finite $\hat{\theta}$. We know from part (1) of Proposition 3 that $\lim_{z \to -\infty} \theta^*(z)$ is finite. Therefore the indifference condition (9) cannot hold for any finite x_I^*.

Claim 6. $\lim_{z \to -\infty} x_I^*(z)$ is not equal to minus infinity.

Suppose to the contrary that $\lim_{z \to -\infty} x_I^*(z)/z = \gamma \geq 0$. First, consider

$$\lim_{z \to -\infty} \Pr[\theta > \hat{\theta}|x_I^*, y_i = 1] \leq \lim_{z \to -\infty} \frac{J(\hat{\theta}, z)\left(1 - \Phi\left(\frac{\hat{\theta} - x_I^*}{\sigma_x}\right)\right)}{\Phi\left(\frac{\overline{x}(z) - x_I^*}{\sqrt{2}\sigma_x}\right) - \Phi\left(\frac{\underline{x}(z) - x_I^*}{\sqrt{2}\sigma_x}\right)}.$$

The term $J(\hat{\theta}, z)$ goes to 0 at the rate $(1 - \sigma_I/\sigma_U)/\sigma_x$. If $1 - \sigma_I/\sigma_U > \gamma$, the denominator goes to 0 at the rate

$$\frac{1 - \frac{\sigma_I}{\sigma_U} - \gamma}{\sqrt{2}\sigma_x} < \frac{1 - \frac{\sigma_I}{\sigma_U}}{\sigma_x}.$$

Therefore the ratio goes to 0 as z goes to minus infinity. If $1 - \sigma_I/\sigma_U \leq \gamma$, the denominator does not vanish. So the ratio again goes to zero. Similar reasoning establishes that $\lim_{z \to -\infty} \Pr[\theta > \hat{\theta}|z, x_I^*, y_i = 1, z \sim I] = 0$. By Claim 2, we therefore have

$$\lim_{z \to -\infty} P(\hat{\theta}|z, x_I^*, 1) = 1 > c$$

if $\lim_{z \to -\infty} x_I^*(z) = -\infty$. Again this contradicts the indifference condition (9) for equilibrium.

Claims 5 and 6 together imply that $\lim_{z \to -\infty} x_I^*(z) = +\infty$. ■

Lemma 2. At $z = z' = \tilde{z}$ and $\theta = \theta'$, the cut-off types who are indifferent between attacking and not attacking satisfy:

$$\frac{\partial \hat{x}_I}{\partial \theta} > \frac{\partial \hat{x}_m}{\partial \theta} > \frac{\partial \hat{x}_U}{\partial \theta} > 0.$$

Proof. We proceed in a number of steps.

Claim 1. $\Pr[y_i = 1|z, x_i]$ is increasing in x_i for $x_i < z$ and decreasing in x_i for $x_i > z$.

Using the definition of T_2 and T_4 in the proof of Claim 2 in Lemma 1, we have

$\Pr[y_i = 1|z, x_i] = wT_2 + wT_4$. Take derivative with respect to x_i to get:

$$\begin{aligned}\frac{\partial \Pr[y_i = 1|z, x_i]}{\partial x_i} &= -w\frac{\beta}{\sqrt{1+\beta\sigma_x}}\left[\phi\left(\frac{\overline{x}(z) - X_i}{\sqrt{1+\beta\sigma_x}}\right) - \phi\left(\frac{\underline{x}(z) - X_i}{\sqrt{1+\beta\sigma_x}}\right)\right] \\ &\quad - (1-w)\frac{1}{\sqrt{2}\sigma_x}\left[\phi\left(\frac{\overline{x}(z) - x_i}{\sqrt{2}\sigma_x}\right) - \phi\left(\frac{\underline{x}(z) - x_i}{\sqrt{2}\sigma_x}\right)\right] + \frac{\partial w}{\partial x_i}F(\beta, z, x_i)\end{aligned}$$

where

$$F(\beta, z, x_i) = \Phi\left(\frac{\overline{x}(z) - X_i}{\sqrt{1+\beta\sigma_x}}\right) - \Phi\left(\frac{\underline{x}(z) - X_i}{\sqrt{1+\beta\sigma_x}}\right) - \Phi\left(\frac{\overline{x}(z) - x_i}{\sqrt{2}\sigma_x}\right) + \Phi\left(\frac{\underline{x}(z) - x_i}{\sqrt{2}\sigma_x}\right),$$

and $X_i = \beta x_i + (1-\beta)z$. It is straightforward to show that $F(\beta, z, x_i)$ is decreasing in β, with $F(1, z, x_i) = 0$. Thus, $F(\beta, z, x_i) > 0$ for $\beta < 1$. Since $\overline{x}(z) + \underline{x}(z) = 2z$, the first two terms are positive if $x_i < z$. Since $\partial w/\partial x_i > 0$ for $x_i < z$, the third term is positive as well. Therefore, the derivative is positive. If $x_i > z$, then the opposite is true.

Claim 2. When $c = 0.5$,

$$\frac{\partial \hat{x}_I(\theta', z')}{\partial \theta} = \frac{-p(\theta'|z', x')}{\int_{-\infty}^{\theta'} \frac{J(\theta,z')}{J(\theta',z')}\frac{\partial p(\theta|z',x')}{\partial x}\,\mathrm{d}\theta} \quad \text{and} \quad \frac{\partial \hat{x}_U(\theta', z')}{\partial \theta} = \frac{-p(\theta'|z', x')}{\int_{-\infty}^{\theta'} \frac{1-J(\theta,z')}{1-J(\theta',z')}\frac{\partial p(\theta|z',x')}{\partial x}\,\mathrm{d}\theta}.$$

When $c = 0.5$, the value of x' that satisfies $P(\theta'|z', x', 1) = c$ is $x' = z'$. By Claim 1, $\partial \Pr[y_i = 1|z', x']/\partial x_i = 0$ at this point. Since

$$P(\theta'|z', x', 1) = \frac{\int_{-\infty}^{\theta'} J(\theta, z')p(\theta|z', x')\,\mathrm{d}\theta}{\Pr[y_i = 1|z', x']},$$

the claim follows by the implicit function theorem. The expression for $\partial x_U/\partial\theta$ is derived in a similar fashion.

Claim 3. When $c = 0.5$,

$$\int_{-\infty}^{\theta'} \frac{J(\theta, z')}{J(\theta', z')}\frac{\partial p(\theta|z', x')}{\partial x}\,\mathrm{d}\theta > \int_{-\infty}^{\theta'} \frac{\partial p(\theta|z', x')}{\partial x}\,\mathrm{d}\theta > \int_{-\infty}^{\theta'} \frac{1 - J(\theta, z')}{1 - J(\theta', z')}\frac{\partial p(\theta|z', x')}{\partial x}\,\mathrm{d}\theta.$$

Using integration-by-parts, the expression on the left-hand-side can be written as

$$\frac{\partial P(\theta'|z', x')}{\partial x} - \frac{1}{J(\theta', z')}\int_{-\infty}^{\theta'} \frac{\partial J(\theta, z')}{\partial \theta}\frac{\partial P(\theta|z', x')}{\partial x}\,\mathrm{d}\theta.$$

The function $J(\theta, z')$ is increasing in θ for all $\theta < z'$. When $c = 0.5$, $\theta' = z'$. Therefore $\partial J/\partial\theta > 0$ for $\theta < \theta'$. Furthermore, since an increase in x leads to a first-order stochastic increase in the posterior distribution of θ, we have $\partial P/\partial x < 0$ for all θ. Thus,

the first inequality of the claim follows. The second inequality can be established in a similar way.

Since x' satisfies the indifference condition $P(\theta'|z', x') = c$ of the "mute model," by the implicit function theorem we obtain:

$$\frac{\partial \hat{x}_m(\theta', z')}{\partial \theta} = \frac{-p(\theta'|z', x')}{\int_{-\infty}^{\theta'} \frac{\partial p(\theta|z', x')}{\partial x} \, \mathrm{d}\theta}.$$

The ranking of the partial derivatives in the lemma then follows by Claims 2 and 3. Finally, since $P(\theta|z, x, 0)$ increases in θ and decreases in x, we have $\partial \hat{x}_U / \partial \theta > 0$. ■

Lemma 3. At $z = z' = \tilde{z}$ and $\theta = \theta'$,

$$\frac{\partial \hat{x}_I}{\partial z} + \frac{\partial \hat{x}_I}{\partial \theta} = \frac{\partial \hat{x}_m}{\partial z} + \frac{\partial \hat{x}_m}{\partial \theta} = \frac{\partial \hat{x}_U}{\partial z} + \frac{\partial \hat{x}_U}{\partial \theta} = 1$$

Proof. Write the relevant indifference conditions in the following form:

$$\begin{aligned}
\tau_m(\theta, z, x) &\equiv P(\theta|z, x) - c = 0, \\
\tau_I(\theta, z, x) &\equiv \int_{-\infty}^{\theta} J(t, z) p(t|z, x) \, \mathrm{d}t - c \Pr[y_i = 1|z, x] = 0, \\
\tau_U(\theta, z, x) &\equiv \int_{-\infty}^{\theta} (1 - J(t, z)) p(t|z, x) \, \mathrm{d}t - c(1 - \Pr[y_i = 1|z, x]) = 0.
\end{aligned}$$

We prove the lemma in a number of steps.

Claim 1. At $z = z' = \tilde{z}$ and $\theta = \theta'$, $\partial \hat{x}_m / \partial z + \partial \hat{x}_m / \partial \theta = 1$.

In the "mute model," We have

$$\begin{aligned}
\frac{\partial \tau_m(\theta', z', x')}{\partial z} &= -w \frac{1-\beta}{\sqrt{\beta}\sigma_x} \phi\left(\frac{\theta' - X'}{\sqrt{\beta}\sigma_x}\right) + \frac{\partial w}{\partial z}\left(\Phi\left(\frac{\theta' - X'}{\sqrt{\beta}\sigma_x}\right) - \Phi\left(\frac{\theta' - x'}{\sigma_x}\right)\right) \\
&= -w \frac{1-\beta}{\sqrt{\beta}\sigma_x} \phi\left(\frac{\theta' - X'}{\sqrt{\beta}\sigma_x}\right),
\end{aligned}$$

where the last equality follows because $X' = \beta x' + (1 - \beta) z' = x'$. Similarly,

$$\frac{\partial \tau_m(\theta', z', x')}{\partial x} = -w \frac{\beta}{\sqrt{\beta}\sigma_x} \phi\left(\frac{\theta' - X'}{\sqrt{\beta}\sigma_x}\right) - (1 - w) \frac{1}{\sigma_x} \phi\left(\frac{\theta' - x'}{\sigma_x}\right)$$

It is straightforward to see that at the point (θ', z', x'),

$$\frac{\partial \tau_m}{\partial z} + \frac{\partial \tau_m}{\partial x} = -\frac{\partial \tau_m}{\partial \theta}.$$

The claim then follows by the implicit function theorem.

Claim 2. At $z = z' = \tilde{z}$ and $\theta = \theta'$,

$$\begin{aligned}
\frac{\partial \tau_I}{\partial x} &= J(\theta', z')\frac{\partial P(\theta'|z', x')}{\partial x} - D, \\
\frac{\partial \tau_U}{\partial x} &= (1 - J(\theta', z'))\frac{\partial P(\theta'|z', x')}{\partial x} + D;
\end{aligned}$$

with $D < 0$.

First, note that $\partial \Pr[y_i = 1|z', x']/\partial x = 0$ by Claim 1 of Lemma 2. Rewrite τ_I using integration-by-parts and then take derivative with respect to x to get

$$D = \int_{-\infty}^{\theta'} \frac{\partial J(\theta, z')}{\partial \theta}\frac{\partial P(\theta|z', x')}{\partial x}\, \mathrm{d}\theta.$$

This term is negative because $\partial J/\partial\theta > 0$ for $\theta < \theta'$ and $\partial P/\partial x < 0$. The derivation of $\partial \tau_U/\partial x$ follows the same lines.

Claim 3. At $z = z' = \tilde{z}$ and $\theta = \theta'$,

$$\begin{aligned}
\frac{\partial \tau_I}{\partial z} &= J(\theta', z')\frac{\partial P(\theta'|z', x')}{\partial z} + Q, \\
\frac{\partial \tau_U}{\partial z} &= (1 - J(\theta', z'))\frac{\partial P(\theta'|z', x')}{\partial z} - Q.
\end{aligned}$$

where $Q = D$.

We can write

$$\tau_I(\theta, z, x) = w(T_1 - cT_2) + (1 - w)(T_3 - cT_4).$$

Therefore,

$$\frac{\partial \tau_I(\theta', z', x')}{\partial z} = w\frac{\partial}{\partial z}(T_1 - cT_2) + (1 - w)\frac{\partial}{\partial z}(T_3 - cT_4),$$

with a term involving $\partial w/\partial z$ that vanishes because, from the proof of part (3) of Proposition 3, $T_1 - cT_2 = T_3 - cT_4 = 0$ at the point (θ', z', x') if $z' = \tilde{z}$.

Consider the derivative of the term $T_1 - cT_2$.

$$\frac{\partial}{\partial z}(T_1 - cT_2) = \int_{-\infty}^{\theta'} J(t, z') \frac{1}{\sqrt{\beta}\sigma_x} \frac{\partial \phi\left(\frac{t - X'}{\sqrt{\beta}\sigma_x}\right)}{\partial z} \, \mathrm{d}t + \int_{-\infty}^{\theta'} \frac{\partial J(t, z')}{\partial z} \frac{1}{\sqrt{\beta}\sigma_x} \phi\left(\frac{t - X'}{\sqrt{\beta}\sigma_x}\right) \mathrm{d}t$$
$$- c\left[\frac{\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\overline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right) - \frac{-\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\underline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right)\right].$$

Use integration-by-parts on the first term to get

$$\begin{aligned}
\frac{\partial}{\partial z}(T_1 - cT_2) = {} & \frac{1}{w} J(\theta', z') \frac{\partial P(\theta' | z', x')}{\partial z} \\
& + \int_{-\infty}^{\theta'} \left((1-\beta)\frac{\partial J(t, z')}{\partial \theta} + \frac{\partial J(t, z')}{\partial z}\right) \frac{1}{\sqrt{\beta}\sigma_x} \phi\left(\frac{t - X'}{\sqrt{\beta}\sigma_x}\right) \mathrm{d}t \\
& - c\left[\frac{\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\overline{x}(z) - X'}{\sqrt{1+\beta}\sigma_x}\right) - \frac{-\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\underline{x}(z) - X'}{\sqrt{1+\beta}\sigma_x}\right)\right].
\end{aligned}$$

From this, we obtain:

$$\begin{aligned}
\frac{\partial}{\partial z}(T_1 - cT_2) = {} & \frac{1}{w} J(\theta', z') \frac{\partial P(\theta' | z', x')}{\partial z} \\
& + \frac{\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\overline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right) \Phi\left(\frac{\theta' - \frac{X' + \beta\overline{x}(z')}{1+\beta}}{\sqrt{\frac{\beta}{1+\beta}}\sigma_x}\right) \\
& - \frac{-\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\underline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right) \Phi\left(\frac{\theta' - \frac{X' + \beta\underline{x}(z')}{1+\beta}}{\sqrt{\frac{\beta}{1+\beta}}\sigma_x}\right) \\
& - c\left[\frac{\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\overline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right) - \frac{-\frac{\mathrm{d}\kappa}{\mathrm{d}z} + \beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\underline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right)\right] \\
= {} & \frac{1}{w} J(\theta', z') \frac{\partial P(\theta' | z', x')}{\partial z} \\
& + \frac{\beta}{\sqrt{1+\beta}\sigma_x} \phi\left(\frac{\overline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right) \left[\Phi\left(\frac{\theta' - \frac{X' + \beta\overline{x}(z')}{1+\beta}}{\sqrt{\frac{\beta}{1+\beta}}\sigma_x}\right) - \Phi\left(\frac{\theta' - \frac{X' + \beta\underline{x}(z')}{1+\beta}}{\sqrt{\frac{\beta}{1+\beta}}\sigma_x}\right)\right],
\end{aligned}$$

where the second equality uses the fact that $z' = \theta' = X'$ and $c = 0.5$. Similarly,

$$
\begin{aligned}
\frac{\partial}{\partial z}(T_3 - cT_4) &= \frac{1 + \frac{\mathrm{d}\kappa}{\mathrm{d}z}}{\sqrt{2}\sigma_x}\phi\left(\frac{\overline{x}(z') - x'}{\sqrt{2}\sigma_x}\right)\Phi\left(\frac{\theta' - \frac{x' + \overline{x}(z')}{2}}{\sqrt{\frac{1}{2}}\sigma_x}\right) \\
&\quad - \frac{1 - \frac{\mathrm{d}\kappa}{\mathrm{d}z}}{\sqrt{2}\sigma_x}\phi\left(\frac{\underline{x}(z') - x'}{\sqrt{2}\sigma_x}\right)\Phi\left(\frac{\theta' - \frac{x' + \underline{x}(z')}{2}}{\sqrt{\frac{1}{2}}\sigma_x}\right) \\
&\quad - c\left[\frac{1 + \frac{\mathrm{d}\kappa}{\mathrm{d}z}}{\sqrt{2}\sigma_x}\phi\left(\frac{\overline{x}(z') - x'}{\sqrt{2}\sigma_x}\right) - \frac{1 - \frac{\mathrm{d}\kappa}{\mathrm{d}z}}{\sqrt{2}\sigma_x}\phi\left(\frac{\underline{x}(z') - x'}{\sqrt{2}\sigma_x}\right)\right] \\
&= \frac{1}{\sqrt{2}\sigma_x}\phi\left(\frac{\overline{x}(z') - x'}{\sqrt{2}\sigma_x}\right)\left[\Phi\left(\frac{\theta' - \frac{x' + \beta\overline{x}(z')}{2}}{\sqrt{\frac{1}{2}}\sigma_x}\right) - \Phi\left(\frac{\theta' - \frac{x' + \beta\underline{x}(z')}{2}}{\sqrt{\frac{1}{2}}\sigma_x}\right)\right].
\end{aligned}
$$

Combining the two terms, and noting that

$$
\begin{aligned}
Q &= \frac{w\beta}{\sqrt{1+\beta}\sigma_x}\phi\left(\frac{\overline{x}(z') - X'}{\sqrt{1+\beta}\sigma_x}\right)\left[\Phi\left(\frac{\theta' - \frac{X' + \beta\overline{x}(z')}{1+\beta}}{\sqrt{\frac{\beta}{1+\beta}}\sigma_x}\right) - \Phi\left(\frac{\theta' - \frac{X' + \beta\underline{x}(z')}{1+\beta}}{\sqrt{\frac{\beta}{1+\beta}}\sigma_x}\right)\right] \\
&\quad + \frac{(1-w)}{\sqrt{2}\sigma_x}\phi\left(\frac{\overline{x}(z') - x'}{\sqrt{2}\sigma_x}\right)\left[\Phi\left(\frac{\theta' - \frac{x' + \beta\overline{x}(z')}{2}}{\sqrt{\frac{1}{2}}\sigma_x}\right) - \Phi\left(\frac{\theta' - \frac{x' + \beta\underline{x}(z')}{2}}{\sqrt{\frac{1}{2}}\sigma_x}\right)\right] \\
&= D,
\end{aligned}
$$

we obtain $\partial\tau_I/\partial z = J\partial P/\partial z + D$. Moreover, since $\tau_U = \tau_m - \tau_I$, this implies $\partial\tau_U/\partial z = (1 - J)\partial P/\partial z - D$.

Claims 2 and 3 imply that

$$
\frac{\partial\tau_I}{\partial x} + \frac{\partial\tau_I}{\partial z} = J\left(\frac{\partial P(\theta'|z', x')}{\partial x} + \frac{\partial P(\theta'|z', x')}{\partial z}\right).
$$

From Claim 1, the term in parenthesis is equal to $-p(\theta'|z', x')$. Therefore, $\partial\tau_I/\partial x + \partial\tau_I/\partial z = -\partial\tau_I/\partial\theta$. By the implicit function theorem, the lemma follows. ■