Optimal (Non-)Repression Under Preference Falsification

Ming-yen Ho *

October 10, 2020

I analyze a dictator's choice of repression level in face of an unexpected shock to regime popularity. In the model citizens could falsify their preferences to feign support of whichever political camp that appears popular. Repression creates fear but also provokes anger and makes the moderates sympathize with the opposition. The dictator considers the tradeoffs and decides the repression level that maximizes regime support in light of his information. Depending on model parameters, the repression level results in regime survival, civil war, voluntary power sharing, or a dictatorship by the opposition. A regime that miscalculates could repress too much or little, leading to outcomes worse than optimal. The model thus provides a general explanation of various outcomes of revolutions and protests observed in history.

1 Introduction

Regime changes such as the Eastern European Revolutions in 1989 occur often suddenly and unexpectedly (Kuran, 1991). Non-revolutions can also be surprising: regimes that have proved terribly inefficient and unpopular such as North Korea have appeared stable for decades. In the Arab Spring, one of the more recent example of surprising revolutions, there were successful revolutions (Tunisia and, temporarily, Egypt) and could-be revolutions that developed into protracted civil wars (Syria). When faced with sudden protests, some dictatorships that ignored the initial signs of discontent were quickly toppled, such as the Pahlavi Dynasty of Iran and the Communist Party in East Germany (Kuran, 1989; Lohmann, 1994). Some regimes tried to suppress fledgling protests, which served only to ignite greater backlash, leading to the regimes' eventually downfall, for example Yanukovych's regime in Ukraine. On the other hand, the Chinese Communist Party harshly repressed widespread demonstrations in 1989 and then enjoyed ruling over a generally stable society for decades. After violently suppressing a student movement in 1980, the South Korean junta allowed for democratic elections in 1987.

What could be the commonalities underneath such wide variety of outcomes? Kuran (1989, 1995) argues that studies of regime stability and transitions overlook the interdependencies of individual decisions: that a person privately detests the regime is not a sufficient reason for her to join a protest. If she expects the regime to be popular and the opposition is weak, social pressures will force her to remain silent and appear content or even supportive. The person protests only if she expects other individuals to join too. However,

*2nd year MA Analytical Political Economy student, Departments of Economics and Political Science, Duke University (mh504@duke.edu). I am especially grateful to Professors Charles Becker, Timur Kuran, and Edward Tower for their mentorship and thoughtful suggestions.

such widespread "preference falsification" implies that massive social discontent could be latent as long as people choose to lie. At the same time, a sudden shock to the regime's appeared stability such as a small-scale protest could attract individuals on the brink of protesting into the fray. This in turn could cause other onlookers to jump on the "revolutionary bandwagon", transforming an initially small small protest into a huge revolution.

Given preference falsification and the interdependencies of people's actions, authoritarian regimes are always vulnerable to sudden shocks and seemingly small protests. This gives dictators a good reason to suppress minor protests forcefully before they materialize into an irresistible revolutionary wave. Nevertheless, heavy repression may backfire either by inducing bystanders' anger (Aytaç, Schiumerini, & Stokes, 2018; Opp & Roehl, 1990) or by informing people uninterested in politics about the repressive "type" of the regime, mobilizing them against the government (Lohmann, 1994). A sophisticated autocrat must balance between the chance that an initially minor shock could snowball into a preventable revolution and the risk of inciting further dissent by repression. However, this is likely to be a difficult task given that the limited information the regime has on people's private preferences.

I study how a dictator chooses an optimal level of repression in face of an unexpected shock to regime stability, knowing that citizens engage in preference falsification and that repression may provoke both the opposition's anger and the moderates' sympathy with the opposition. By constructing a citizen's utility as a linear function of her public regime preference and integrity loss due to deviation from private preference, I am able to derive thresholds under which citizens with different private preferences change their public preferences according to expected support of the regime and the opposition. A dictator fully or partially informed about the parameters of these thresholds attempts to exploit them in choosing repression that maximizes expected regime support.

A perfectly informed dictator will not repress as much as an uninformed dictator and will also obtain better outcomes. In particular, if the shock is weak, it will dissipate and should be left to its own. If the shock is moderate and the opposition is strong enough, the dictator does not repress and voluntarily democratizes. If the anger provoked by the repression is strong but the opposition's real strength is weak, the moderates will side with the regime in a civil war with the opposition. However, if the shock is large or if emotions from observing repression is weak, the dictator almost always represses. A less informed dictator can "overshoot" or "undershoot" his repression level and lose power. In particular, when the shock is low, excessive optimism about the regime's stability makes a dictator more likely to "undershoot" repression, allowing the opportunity for the shock to snowball into a revolution of a larger scale. If the shock is large, however, the optimistic dictator may "overshoot" his repression while a better alternative may be to share power and voluntarily democratize. To contrast, a "paranoid" dictator who underestimates his regime's

stability may repress too much if the shock is small, mobilizing anger from the opposition. If the shock is large, however, the "paranoid" dictator may concede when more repression is optimal. Thus, the observed wide range of outcomes in regimes' decisions to repress or not repress can be explained by authoritarian regimes' inability to fully understand their citizens' private preferences before reacting to a crisis.

2 Protest Thresholds and Repression

Social scientists have long recognized that people's decisions to join a collective action may depend on the participation of others (Granovetter, 1978; Schelling, 1978). This is especially true for participating in political protests against a dictatorship capable of suppressing dissent without restraint. One view is that this presents a collective action problem in which individuals could free-ride on others' costly decisions to participate, making protests a game of "strategic substitutes" (Cantoni, Yang, Yuchtman, & Zhang, 2019; Tullock, 1971). The standard "strategic complementarity" assumption states instead that since a protest's probability of success and therefore its expected reward increases with the number of its participants, greater participation makes joining the protest more attractive (DeNardo, 1985; Morris & Shin, 2002; Oberschall, 1994; Passarelli & Tabellini, 2017). However, the number of willing participants of a potential revolution, which is crucial to each individual's calculus in deciding whether to join a fomenting protest, is ex ante uncertain, creating a coordination problem. This cultivates a fertile ground for a literature on how governments, opposition parties, and citizens could strategically send signals through their actions to manipulate beliefs as to the level of anti-regime sentiment and regime strength, thereby influencing the outcome of the revolution (Bueno de Mesquita, 2010; Chwe, 2000; Edmond, 2013; Ginkel & Smith, 1999).

Often the citizen's problem lies not just in the uncertainty of other citizens' actions; i.e., the probability of success itself, but also uncertainty in the actual merit, i.e. payoff, of a successful revolution. Under a regime that restricts and manipulates information, citizens will have difficulty comparing the current regime and alternative political arrangements espoused by the opposition (Shadmehr & Bernhardt, 2011). Lohmann (1994) shows that participation in protests by other citizens conveys information to the politically moderate and uninformed about the likely type of the regime and the merits of protesting. An unexpected protest turnout may in turn induce the moderates and the uninformed to join the protest in later stages, leading to an "information cascade" that allows initially small protests to snowball into a popular revolution (Bikchandani, Hirshleifer, & Welsh, 1992). This explains why revolutions are rare but when they appear, they come in waves and spread easily to other countries (Chen & Suen, 2016).

A key implication of these "cascade models" is that revolutions are often surprising and unpredictable due to the rational individual's tendency to mask their true political prefer-

ences when they believe that their private opinion is unpopular (Kuran, 1989). A flip side of the coin is that authoritarian regimes cannot gauge how susceptible to shocks they truly are since they are almost always ostensibly popular. Knowing this, a rational dictatorship will attempt to strengthen its dissenters' incentive to lie by influencing citizens' beliefs as to the costs of revolting and the regime's infallibility. Edmond (2013) considers how differences in information technology affect the ability of autocracies to control information and enhance its chance of survival. If the information technology is centralized and the marginal cost of controlling additional signal source is decreasing with number of signals (e.g. mass media), it is easier for the regime to manipulate citizens' beliefs and prevent a revolution. Ellis and Fender (2010) argue that poorer information flows encourage the ruling elites to maintain oppressive systems of oligarchy and tolerate a positive probability of revolution, since the chance of dissatisfied citizens coordinating on a revolution is low.

To prevent minor incidents from spiralling into widespread revolt, the regime may repress open dissenters to avoid them from signalling the regime's unpopularity to citizens at the margin of protesting. An alternative may be to appease dissenters by accommodating their demands. However, Ginkel and Smith (1999) suggest that regimes whose actual strength is imperfectly known rarely make compromises in crisis, since offering accommodation instead of repressing signals weakness and actually encourages protests. Another similar paradox is that the more repressive the regime, the dissidents' decision to mobilize protests makes their claims as to the government's vulnerability more credible. Under high repression, revolutions and protests will be rarer but once they occur, they are more likely to be joined by the bystanders and be successful. Rubin (2014) argues that regimes with centralized power to impose sanctions are able to suppress small external shocks easily, but greater coercion makes preference falsification more prevalent. This enables more serious shocks to trigger large-scale revolutionary cascades that could dramatically change the political equilibrium suddenly and unexpectedly.

Empirically, the relationship between the level of government repression and the intensity of political protests is inconclusive, suggesting that repression presents trade-offs and is often counterproductive (Opp & Roehl, 1990; Pierskalla, 2010; Shadmehr & Bernhardt, 2011; Siegel, 2011). Opp and Roehl (1990) contend that repression has a direct effect in reducing protest turnout by increasing cost of dissent, but it may facilitate mobilization of the opposition indirectly. Repression perceived as illegitimate provokes emotions such as anger towards the regime and sympathy to the opposition that increases the social incentives to participating in the opposition. Furthermore, the use of violence by the government may encourage the opposition to respond with violence, an option now perceived as legitimate and necessary. Pierskalla (2010) suggests under a pure two-player setting a rational regime would compromise in lieu of repressing if its strength is too weak and a rational opposition would not launch an open protest if the regime is perceived as too strong. However, if third parties capable of determining outcomes like the military

are present, both the government and the opposition may escalate violence to signal their strength to the third parties.

Not only does the question of whether or not to repress present tradeoffs, but also it also affects the *method* of repression. Shadmehr (2015) considers a model under which a revolutionary entrepreneur chooses his agenda, citizens with heterogeneous preference choose their effort of supporting the revolution, and the regime decides its policy of repression. The results suggest that increasing the minimum level of punishment deters moderate citizens from participating and radicalizes the optimal revolutionary agenda. If the regime imposes indiscriminate punishment to all citizens, revolution is likely but its agenda would be more moderate.

The literature on the strategic complementarity of protests generally analyses how various actors take strategic actions to facilitate or block the coordination of protests. The citizens' decisions are simultaneously made and revealed, thereby immediately deciding the outcome: either the regime is toppled if participation passes a given threshold, or else the regime survives (Bueno de Mesquita, 2010; Edmond, 2013; Ginkel & Smith, 1999; Shadmehr & Bernhardt, 2011). The downside of these one-shot game approaches is that they can only analyse the actors' strategic calculations *prior* to the protests' occurrence or non-occurrence, which is largely determined by given parameters. This restricts the study of how protests and revolutions could evolve across time as actors make decisions based on information from sequential events, generating interesting phenomena such as bandwagoning or free-riding.

The "informational cascades" approach instead focuses on protests/revolutions as a path-dependent *process* in which individuals decide whether or not participate sequentially in time. Each individual takes into account the actions of preceding individuals who have acted before deciding to revolt or not (Bikchandani et al., 1992; Chen & Suen, 2016; Lohmann, 1994). These models capture the elements of path-dependence and randomness inherent in the process of protests and revolutions in which individuals make their choices as events unfold. However, they generally study how individuals decide and/or how revolutionary groups could exploit cascades to generate revolutions, but not how the regime may counteract to stem a developing revolutionary bandwagon. Similarly, the literature on "preference falsification" investigates how various distributions of private preferences and protest thresholds or how different sociopolitical institutions affect the likelihood and magnitude of sudden revolutions (Kuran, 1989, 1995; Rubin, 2014; Yin, 1998). How rational dictators, knowing that citizens falsify their preferences and that the regime's apparent stability may be fragile, may implement actions to reduce their regimes' susceptibility to sudden shocks remains an unexplored implication.

Existing studies on repression concentrate on explaining the tradeoffs of repression

and suggest conditions under which repression would be successful or counterproductive (Aytaç et al., 2018; Opp & Roehl, 1990; Shadmehr, 2015; Siegel, 2011). Why sophisticated autocrats, who understand that protests may backfire, have nevertheless made fatal mistakes either by repressing "too much" or "too little", has received less attention. In models that do analyze the regime's use of repression or concession in strategic interaction with the oppositions and third parties, citizens are treated as a single unitary actor rather than multiple actors who make interdependent decisions (Ginkel & Smith, 1999; Pierskalla, 2010). This fails to account for important mechanisms such as preference falsification, informational cascades, and strategic complementarity that are crucial in determining protest turnout and success.

In this paper I shall combine insights from the literature on the interdependencies of protest decisions and the dilemma of repression to offer an integrated account of why dictators, understanding the tradeoffs of repression and the logic of preference falsification, may tolerate, repress, or accommodate in face of protests. The paper also analyzes conditions under which the regime chooses to repress even though they know that repression will not be entirely successful, thereby escalating conflict into a civil war. Besley and Persson (2011)'s model predicts that domestic peace, repression (use of violence by the incumbent), and civil war (use of violence by both the incumbent and the opposition) are three equilibrium outcomes *ordered* by domestic economic and political determinants of violence. Specifically, higher income, higher spending on public goods, and cohesive political institutions that constrain power reduce the incentive of both the incumbent and the opposition to capture power by violence, while greater rents from natural resources increase the incentive for fighting. "The threat of revolution" is attributed as a key cause of elites' voluntary democratization in major accounts of democratization (Acemoglu & Robinson, 2006; O'Halloran, Leventoglu, & Epstein, 2012). Indeed, when the tide of revolution is considered irresistable or that suppression is too costly, in this model the dictatorship also voluntarily democratize to preserve part of their rents.

3 Model

3.1 General Setting

Consider a society populated by a continuum of citizens with total measure 1 and ruled by an authoritarian regime Party 0. Let $[0, 1] \in \mathbb{R}$ represent the space of policies or political ideology. Party 0 implements policy 0. There exists also an opposition Party 1 that advocates for an alternative policy regime 1. There are three types of citizens: "Type 0" citizens have 0 as their policy/ideological bliss point, whose proportion among the total population is $\alpha < 1$. "Type 1" citizens have policy/ideological bliss point 1, who take up $\beta < 1$ of the population. There are also "Type 0.5" citizens with policy bliss point 0.5, and their proportion is $1 - \alpha - \beta > 0$. That citizens' policy bliss points do not distribute continuously over $[0, 1]$ is assumed not just because of analytical simplicity, but to reflect actual

informational and structural constraints that make the set of feasible/imaginable political options limited. Only a few (and often only two) parties are able to organize effectively to mobilize sufficient political appeal to a tied bundle of policy programs.

Following Kuran (1989), each citizen i has a private policy preference x_i which they could hide by expressing their public preference y_i. $x_i = 0$ for Type 0 citizens, $x_i = 1$ for Type 1, and $x_i = 0.5$ for Type 0.5. Each citizen's utility function $U(y_i, x_i)$ consists of two components, R, the *reputational utility* function, and I, the *integrity* utility function. By expressing public preference $y_i \in [0, 1]$, the citizen could obtain social and material benefits from the party she publicly supports. In particular, such utility is increasing in the proportion of *expected* support towards the camp of y_i. Let $\hat{\alpha}$ denote the expected support of Party 0 and $\hat{\beta}$ the expected support of Party 1. Note that α, β and therefore $1 - \alpha - \beta$, the actual proportion of the parties' support, is unknown to all citizens and parties precisely due to the possible divergence between y_i and x_i.

We assume that $R(y_i, \hat{\alpha}, \hat{\beta}) = (1 - y_i)\hat{\alpha} + y_i\hat{\beta}$. Then if $y_i = 0$, or the case when the citizen publicly supports Party 0, she expects to obtain utility $\hat{\alpha}$. If she supports Party 1 instead, she obtains utility $\hat{\beta}$. Note that there is some positive utility if the citizen chooses $y_i \in (0, 1)$, but in general it will yield her strictly less utility than either 1 or 0 as long as $\hat{\alpha} \neq \hat{\beta}$. We allow positive utility from choosing $y_i \in (0, 1)$ since supporters of the two parties will try to woo the moderates into their camp by expending some benefits proportional to their parties' ability to deliver reputational utility.

We assume $I(y_i, x_i, \delta) = -\delta|y_i - x_i|$. If $y_i \neq x_i$, the individual suffers some psychological loss from her inability to speak out her true preference. This loss is scaled by the parameter δ, which determines the relative strength between reputation and integrity in the individual's consideration of y_i. For reasons that will be clearer later, δ could also be interpreted as the degree in which the society tolerates and protects free political speech and action. Hence, low δ implies that the individual has more reasons to fear the consequences of expressing unpopular political opinions. I assume that every individual has the same δ in the spirit of this interpretation, implying that every citizen of the same type has the same utility function. Furthermore, following Kuran (1989), each person's weight on public opinion is negligible, so that the individual does not expect her choice of y_i to change $\hat{\alpha}, \hat{\beta}$. Then

$$U(x_i, y_i, \hat{\alpha}, \hat{\beta}, \delta) = R(y_i, \hat{\alpha}, \hat{\beta}) + I(y_i, x_i, \delta) = (1 - y_i)\hat{\alpha} + y_i\hat{\beta} - \delta|y_i - x_i| \tag{1}$$

Given shared expectations $\{\hat{\alpha}, \hat{\beta}\}$ and integrity parameter δ, each type of citizens with private preference x_i simultaneously chooses y_i to maximize U_i. The utility structure is similar to Kuran (1989)'s in that the citizen's utility is the sum of his reputational utility derived from publicly stated preferences and utility derived from integrity, which is proportional to the distance between the citizen's private and public preferences. I differ by

specifically assuming that the reputational gains from publicly supporting a Party is proportional to the Party's expected level of support. I also introduce the integrity parameter δ to study how differences in weights individuals give to integrity affect their propensity to falsify their preferences.

Before diving into the setup of the game, we first characterize the immediate consequences of citizens selecting their public preferences according to such utility functions. In general expectations may not be self-confirming in the sense that given expectations $\{\hat{\alpha}, \hat{\beta}\}$, the actual proportion of citizens expressing $y_i = 0$ and $y_i = 1$ may not equal to $\hat{\alpha}$ and $\hat{\beta}$. Upon observing this outcome, citizens will revise their expectations until an equilibrium is reached.

Definition. *An equilibrium $\{\alpha^*, \beta^*\}$ is a set of expectations such that if $\{\hat{\alpha}, \hat{\beta}\} = \{\alpha^*, \beta^*\}$, then the actual proportion of citizens supporting Party 0 equals α^* and the proportion of citizens supporting Party 1 equals β^*.*

3.2 Possible Equilibria

For given α, β, δ, there could be several possible equilibria since y_i is a function of $\hat{\alpha}$ and $\hat{\beta}$ but not of α, β. We first consider the critical thresholds from which each type of citizens will switch from supporting one y_i to another y_i, which depends on the relative strength of expectations $\hat{\alpha}, \hat{\beta}$, and integrity δ.

Type 0 Citizens. Given $\hat{\alpha}, \hat{\beta}, \delta$, type 0 citizens ($x_i = 0$) choose $y_i = 0$ if $U(y_i = 0|x_i = 0) = (1-y_i)\hat{\alpha} + y_i\hat{\beta} - \delta(y_i - 0)|_{y_i=0} = \hat{\alpha} \geq U(y_i > 0|x_i = 0) = (1-y_i)\hat{\alpha} + y_i\hat{\beta} - \delta y_i$. I assume the tiebreaker convention that if $U(y_i = 0|x_i = 0) \geq U(y_i > 0|x_i = 0)$ for all $y_i > 0$, the individual will follow his private preference and select $y_i = 0$. Rearrange this inequality $\hat{\alpha} \geq (1 - y_i)\hat{\alpha} + y_i\hat{\beta} - \delta y_i$ and we get the condition $\hat{\alpha} \geq \hat{\beta} - \delta$. Type 0 citizen chooses $y_i = 1$ if $U(y_i = 1|x_i = 0) = \hat{\beta} - \delta > (1 - y_i)\hat{\alpha} + y_i\beta - \delta y_i = U(y_i < 1|x_i = 0)$, rearrange and we get the condition $\hat{\alpha} < \hat{\beta} - \delta$.

Comparing the two threshold conditions, Type 0 citizens never choose $y_i \in (0, 1)$. This is driven by the fact that once $\hat{\beta}$ is high enough (and δ low enough) so that choosing $y_i = 0$ is no longer optimal, the linear utility implies that the individual is always better off trading integrity utility at price δ to gain reputational utility with marginal value $\hat{\beta} - \hat{\alpha}$ as implied by the condition $\hat{\alpha} < \hat{\beta} - \delta$.

Type 1 Citizens. For Type 1 citizens the derivation is similar. Comparing $U(y_i = 0|x_i = 1) = \hat{\alpha} - \delta, U(0 < y_i < 1|x_i = 1) = (1 - y_i)\hat{\alpha} + y_i\hat{\beta} - \delta(1 - y_i), U(y_i = 1|x_i = 1) = \hat{\beta}$, we obtain that Type 1 citizens choose $y_i = 0$ if $U(y_i = 0|x_i = 1) = \hat{\alpha} - \delta > \hat{\beta} = U(y_i = 1|x_i = 1)$ and $\hat{\alpha} - \delta > (1 - y_i)\hat{\alpha} + y_i\hat{\beta} - \delta(1 - y_i)$ which yields the condition $\hat{\alpha} > \hat{\beta} + \delta$. Type 1 citizens will publicly support their true preference $y_i = 1$ if $U(x_i = 1|y_i = 1) = \hat{\beta} \geq$

$(1-y_i)\hat{\alpha}+y_i\hat{\beta}-\delta(1-y_i)$, or $\hat{\alpha}\leq\hat{\beta}+\delta$. Again they never choose $y_i\in(0,1)$.

Type 0.5 Citizens. Type 0.5 citizens choose $y_i=0$ if $U(y_i=0|x_i=0.5)=\hat{\alpha}-0.5\delta>(1-y_i)\hat{\alpha}+y_i\hat{\beta}-\delta|y_i-0.5|=U(y_i>0|x_i=0.5)$. Type 0.5 citizens will choose $y_i=0$ only if $\hat{\alpha}>0.5$. When $\hat{\alpha}>0.5$, note that for each $\epsilon>0$, $U(y_i=0.5-\epsilon|x_i=0.5)>U(y_i=0.5+\epsilon|x_i=0.5)$. Intuitively when $\hat{\alpha}>0.5$, or Party 0 is expected to gain majority support, choosing any $y_i=0.5+\epsilon$ yields strictly less reputational utility than the corresponding $y_i=0.5-\epsilon$ that yields the same utility loss from integrity. Then type 0.5 citizens will never consider $y_i>0.5$ if $\hat{\alpha}>0.5$, then we can simply consider $\hat{\alpha}-0.5\delta>(1-y_i)\hat{\alpha}+y_i\hat{\beta}-\delta(0.5-y_i)$ and get $\hat{\alpha}>\hat{\beta}+\delta$.

By a similar argument, Type 0.5 citizens will choose $y_i=1$ if $\hat{\beta}>\hat{\alpha}+\delta$. I claim that in the intermediate range $\left|\hat{\alpha}-\hat{\beta}\right|\leq\delta$, Type 0.5 citizens will chose only $y_i=0.5$. Obviously when $\hat{\alpha}=\hat{\beta}$, $y_i=0.5$ is optimal since any y_i yields the same reputational utility. Suppose $\hat{\beta}>\hat{\alpha}$. By the same argument above, any $y_i>0.5$ is better than corresponding option $y_i<0.5$ that has the same absolute deviation away from 0.5. Consider $U(y_i>0.5|x_i=0.5)-U(y_i=0.5|x_i=0.5)=(1-y_i)\hat{\alpha}+y_i\hat{\beta}-(y_i-0.5)\delta-[0.5(\hat{\alpha}+\hat{\beta})]=(0.5-y_i)(\hat{\alpha}-\hat{\beta}+\delta)$. Since $\hat{\alpha}-\hat{\beta}\geq-\delta$ and $y_i>0.5$, $K\leq0$ with equality when $\hat{\alpha}-\hat{\beta}=-\delta$. Similarly when $\hat{\alpha}>\hat{\beta}$, $U(y_i<0.5|x_i=0.5)-U(y_i=0.5|x_i=0.5)=(1-y_i)\hat{\alpha}+y_i\hat{\beta}-(0.5-y_i)\delta-[0.5(\hat{\alpha}+\hat{\beta})]=(0.5-y_i)(\hat{\alpha}-\hat{\beta}-\delta)\leq0$ since $y_i<0.5$ and $\hat{\alpha}-\hat{\beta}\leq\delta$. This proves that in the intermediate range any $y_i\neq0.5$ will yield less utility than $y_i=0$. Then we have derived the conditions for each type of citizens to choose $y_i=0,1,0.5$:

$$y_i|_{x_i=0}=\begin{cases}0 & \text{if } \hat{\alpha}\geq\hat{\beta}-\delta\\ 1 & \text{if } \hat{\alpha}<\hat{\beta}-\delta\end{cases} \tag{2}$$

$$y_i|_{x_i=0.5}=\begin{cases}0 & \text{if } \hat{\alpha}>\hat{\beta}+\delta\\ 0.5 & \text{if } \left|\hat{\alpha}-\hat{\beta}\right|\leq\delta\\ 1 & \text{if } \hat{\alpha}<\hat{\beta}-\delta\end{cases} \tag{3}$$

$$y_i|_{x_i=1}=\begin{cases}0 & \text{if } \hat{\alpha}>\hat{\beta}+\delta\\ 1 & \text{if } \hat{\alpha}\leq\hat{\beta}+\delta\end{cases} \tag{4}$$

Feasible equilibria. Based on these conditions we can proceed to analyze possible equilibria. First if $\hat{\alpha}>\hat{\beta}+\delta$, all three types of agents will choose $y_i=0$. In this situation the only self confirming equilibrium is $(\hat{\alpha},\hat{\beta})=(1,0)$ since everyone publicly supports incumbent regime Party 1. As long as $\delta\in[0,1),\hat{\alpha}=1>\hat{\beta}+\delta=\delta$, the equilibrium is sustainable. Consider the extreme example where $\{\alpha,\beta\}=\{0,1\}$ but $\{\hat{\alpha},\hat{\beta}\}=\{1,0\}$ and $\delta=0.5$. The society only has Type 1 citizens who privately prefer Party 1. However,

since every individual expects Party 0 to enjoy unanimous support ($\hat{\alpha} = 1$), the reputational utility loss of not supporting 0 is too high compared to the loss of integrity utility in supporting 0. Then every Type 1 citizen takes $y_i = 0$, resulting in everyone publicly supporting the regime and confirming the original expectations. This illustrates how preference falsification could mask the regime's true popularity.

If $\left|\hat{\alpha} - \hat{\beta}\right| \leq \delta$, Type 0 agent chooses $y_i = 0$, Type 0.5 chooses $y_i = 0.5$, Type 1 chooses $y_i = 1$. Then the only equilibrium is $\{\hat{\alpha}, \hat{\beta}\} = \{\alpha, \beta\}$. This is the "truthful equilibrium" where everyone public expresses their private preference. Intuitively, the truthful equilibrium is reached either when both parties are similar enough in expected strength (small $\left|\hat{\alpha} - \hat{\beta}\right|$) or if the integrity parameter δ is high. It is self sustaining as long as $\left|\hat{\alpha} - \hat{\beta}\right| \leq \delta$ continues to hold. Observe that this implies the truthful equilibrium is sustainable *only if* $|\alpha - \beta| \leq \delta$, i.e., the true difference in the support of Parties 0 and 1 must be small enough or the society's tolerance for unpopular opinion is large enough for everyone to continue expressing their genuine political preference without engaging in preference falsfication.

Finally, when $\hat{\alpha} < \hat{\beta} - \delta$, Types 0, 0.5, and 1 all choose $y_i = 1$ and the only self-confirming equilibrium is $\{\hat{\alpha}, \hat{\beta}\} = \{0, 1\}$. We now consider how the size of δ affects the number of possible equilibria. First observe that since $\hat{\alpha}$ and $\hat{\beta}$ is restricted in $[0, 1]$, when $\delta > 1$ the only possible equilibrium is the truthful equilibrium. As discussed, since δ reflects the degree of tolerance and institutional protection of free speech in the society, $\delta > 1$ corresponds to a situation with functional democracy and political freedom. Everyone values speaking one's true belief more than fearing the consequences of expressing unpopular opinion, even when the individual expects disagreement from the overwhelming majority of the society.

Consider $\delta \in [\,|\alpha - \beta|\,, 1)$. Here the truthful equilibrium is feasible since $|\alpha - \beta| \leq \delta$, and all three types of equilibria are possible. If $\delta \in [0, |\alpha - \beta|)$, the truthful equilibrium is not feasible: either people all support Party 0 or Party 1, which depends on which party is expected to possess majority support. This approximates conditions where norms of political tolerance are weak and people expect the incumbent regime to suppress any dissent, leading them to "follow the herd".

3.3 Game Setup

Now we are ready to introduce the game where everyone observes a shock that affects the level of expected support and the incumbent regime Party 0 chooses whether to repress such shocks. The citizens wish to maximize their utility while Party 0 wants to maximize $\hat{\alpha}$ by choosing repression. First, we define shocks as an unforseen event that increases the expected support of Party 1 ($\hat{\beta}$) to the detriment of the expected support of Party 0 ($\hat{\alpha}$). This could be an unexpected large protest, a routine act of brutality by law enforcers of the regime that somehow ignited the anger of onlookers, a sudden call by a respected fig-

ure to resist the regime, or socioeconomic adversities for which the regime is unprepared (Lohmann, 1994; Rubin, 2014). In particular, I assume that the shock manifests itself in the way that it removes s from $\hat{\alpha}$ and transfers it to $\hat{\beta}$; i.e., some subset of the population previously thought to be supporting the regime is now seen to be turning against the regime. The magnitude of the shock is commonly observed by all types of citizens and the parties. The shock is *unexpected* in the sense that it is not chosen by the Party 1.

Returning to the conditions in (1)-(3), we see that for an initial pro-Party 0 equilibrium $\{\hat{\alpha}, \hat{\beta}\} = \{1, 0\}$, a shock of $s < \frac{1-\delta}{2}$ is not enough to make Type 1 and Type 0.5 citizens move away from expressing $y_i = 0$. Then the shock is transient and upon observing nobody choosing $y_i = 1$, the pro-Party 0 equilibrium is restored. If $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, then Type 1 citizens choose $y_i = 1$, Type 0.5 choose $y_i = 0.5$, and Type 0 choose $y_i = 0$. We will end in the truthful equilibrium $\{\hat{\alpha}, \hat{\beta}\} = \{\alpha, \beta\}$. If $s > \frac{1+\delta}{2}$, then every citizen chooses $y_i = 1$. Then an initially small shock may result in larger changes in the observed support of the parties than initially expected. This formalizes the revolutionary cascade model powered by reputational incentives in Kuran (1989). Foreseeing such a domino effect, Party 0 has every incentive to forestall the opposition's support from snowballing.

Upon observing the shock, the incumbent Party 0 could decide whether to repress it by selecting a repression level r. I assume that the regime disregards explicit costs in choosing repression because (1) costs of repression are of second order importance to maintaining appearance of regime support and (2) the regime has no time to increase its investment in coercive forces in reaction to the sudden crisis, and can only use its existing security forces on which it has already committed expenditures. Therefore, I only impose the constraint that if the dictatorship has maximum coercive capability $R < 1$, then $r \in [0, R]$. Furthermore I assume that $R > \frac{1+\delta}{2}$ so that all moderate shocks can be repressed fully. The first-order impact of the repression r is to reduce s: it reduces s by r so that the new expectations given shock s and repression r become $\{\hat{\alpha}_{sr}, \hat{\beta}_{sr}\} = \{\hat{\alpha} - s + r, \hat{\beta} + s - r\}$.

Definition. *A shock $s \leq \hat{\alpha}$ is defined as a change in $\{\hat{\alpha}, \hat{\beta}\}$ that generates a new set of expectations $\{\hat{\alpha}_s, \hat{\beta}_s\} = \{\hat{\alpha} - s, \hat{\beta} + s\}$ shared by all citizens and parties. Before citizens respond to the shock by choosing their public preference y_i, the regime could choose a repression level $r \in [0, \min\{s, R\}]$ such that the expectations is further updated to be $\{\hat{\alpha}_{sr}, \hat{\beta}_{sr}\} = \{\hat{\alpha} - s + r, \hat{\beta} + s - r\}$.*

The citizens then respond to the shock and the repression by choosing y_i. If the citizens observe repression $r > 0$, I introduce two mechanisms that will change their preferences. The *emotional mechanism* applies to Type 1 citizens who already privately resent the regime. Observing positive repression will incite anger and hatred towards Party 0, which manifests itself by increasing the integrity parameter of Type 1 citizens. All else equal, Type 1 citizens are now more willing to choose $y_i = 1$ to openly oppose the regime. I specify $e(r)$ as the function mapping repression level r to increase in Type 1 citizens' integrity.

$e(0) = 0$ and $e'(r) > 0$. Then if Party 0 chooses repression level r, Type 1's utility becomes $(1 - y_i)(\hat{\alpha} - s + r) + y_i(\hat{\beta} + s - r) - (\delta + e(r))(1 - y_i)$.

In addition, there is the *information mechanism* that affects the private preferences of Type 0.5 agents. Repression can deliver a signal to political moderates, informing them that the regime is "bad" and is willing to suppress dissent with brutal force. I assume that for given repression r, for each Type 0.5 agent there is a probability $p(r)$ such that the Type 0.5 citizen upon observing such repression will sympathize with Party 1 and become a Type 1 citizen. With probability $1 - p(r)$ the citizen will remain a Type 0.5 citizen. Then if Party 0 chooses r, then the actual proportion of Type 0 and Type 1 citizens become $\{\alpha, \beta + p(r)(1 - \alpha - \beta)\}$. $p(r)$ is increasing in r and $p(0) = 0$, I shall later assume $p(r) = r$ for notational simplicity but it shall not affect the results. I assume that repression does not affect the preferences of Type 0 agents, but modifying it to benefit Party 1 will strengthen the results.

There is also the implicit *fear mechanism* which lies in how r reduces the impact of shocks by increasing $\hat{\alpha}$ and decreasing $\hat{\beta}$. By suppressing those that participate in the event creating the shock, the regime attempts to maintain image of regime strength and make potential dissenters think twice about the consequences of revealing their true colors. If Type 1 agents, now with integrity parameter $\delta + e(r)$ and comprising $\beta + (1 - \alpha - \beta)r$ of the population, choose $y_i = 0$ upon observing the revised expected support of both camps, repression level r is effective in sustaining high expected regime support $\hat{\alpha}$, the real factor determining regime survival. A dictator who understands the consequences of repression on citizens' integrity utility will balance between the fear mechanism's effect on reputational utility and the emotional and the informational mechanisms' impact on integrity utility.

We consider a society with static equilibrium $\{\hat{\alpha}, \hat{\beta}\} = \{1, 0\}$ and $\delta \in [\,|\alpha - \beta|\,, 1)$, so that all three equilibria are possible. I now introduce the sequence of play:

1. A shock $s \in (0, \hat{\alpha})$ is drawn;
2. Party 0 chooses repression level $r \in [0, \min\{s, R\}]$;
3. Upon observing repression r, all agents update utility functions according to the new expectations $\{1 - s + r, s - r\}$, a proportion r of $1 - \alpha - \beta$ Type 0.5 citizens now prefers $x_i = 1$, and Type 1 citizens (now taking $\beta + r(1 - \alpha - \beta)$ of the population) update their integrity parameter to be $\delta + e(r)$;
4. all types of citizens choose y_i simultaneously to maximize their utility $U = R + I$;
5. y_i is revealed, and new expectations $\{\hat{\alpha}', \hat{\beta}'\}$ are generated by the *actual* proportion of citizens choosing $y_i = 0$ and $y_i = 1$ respectively;

6. After the first sequence ends, we will check if $\{\hat{\alpha}', \hat{\beta}'\}$ is an equilibrium in the sense that $\{\hat{\alpha}', \hat{\beta}'\} = \{1 - s + r, s - r\}$, the expectations that citizens relied on in making their decisions. If there are discrepancies between the two, the citizens will further update their utilities until an equilibrium is reached.

Notice that the limit to the regime's coercive capacity R puts a natural limit to the emotional, informational, and fear mechanisms. In this manner we can analyze how an initially minor shock, amplified by reactions to government repression, may result in a cascade that dramatically changes the political landscape.

4 Analysis

We have not specified how the regime decides level of repression r. The tradeoffs from higher repression, that it increases the integrity of Type 1 citizens and causes some Type 0.5 citizens to become Type 1, probably are not fully known by the regime. Party 0's decision depends on its information set: whether it recognizes that $\{\hat{\alpha}, \hat{\beta}\}$ may not correspond to the actual private preferences of its citizens as well as the values of δ, $p(r)$, and $e(r)$.

I consider three cases: The first and the simplest is the "naive dictator" case where Party 0 believes that the observed $\{\hat{\alpha}, \hat{\beta}\}$ corresponds to actual regime support. In other words, the naive dictator believes his people's public preferences are genuine: only those open defectors who produce shock s support the opposition. Then by choosing repression $r = s$ the dictator believes he can increase regime support without inciting anger or making closet moderates switch to the opposition, since all remaining citizens are loyal supporters of the regime. Second, we explore the "perfect information" case where Party 0 knows $\alpha, \beta, \delta, e(r)$ and $p(r) = r$. The strategy adopted by a perfectly informed dictator is optimal in the sense that it results in the largest possible $\hat{\alpha}$ under various given scenarios of shock levels and parameters. Finally, there is the "imperfect information" case where Party 0 has some estimates of $\{\alpha, \beta\} = \{\bar{\alpha}, \bar{\beta}\}$, $\delta = \bar{\delta}$, $e(r) = \bar{e}(r)$ and $p(r) = \bar{r}$. This situation accords with probably how most sophisticated dictatorships manage a crisis. I analyze primarily the "optimistic" case where $\bar{\alpha} > \alpha$, $\bar{\beta} < \beta$, $\bar{\delta} < \delta$, $\bar{e}(r) < e(r)$ for all r, and $\bar{p}(r) < r$. It is optimistic in the sense that Party 0's estimates of these parameters suggest that the regime is stronger than it truly is. This assumption is natural since preference falsification among the populace will provide an inevitable positive bias to the popularity of the regime. The contrast is the "pessimistic" or the "paranoid" case where the dictator's $\bar{\alpha} < \alpha$, $\bar{\beta} > \beta$, $\bar{\delta} > \delta$, $\bar{e}(r) > e(r)$ for all r, and $\bar{p}(r) > r$.

4.1 The Naive Dictator

Since repression is costless for the naive dictator who wishes to maximize $\hat{\alpha}$, then given any shock s, Party 0 chooses $r = \min\{s, R\}$. Assume that $r = s \leq R$. Given expectations $\{1 - s + r, 1 - r\} = \{1, 0\}$, Type 1 citizens will choose $y_i = 0$ if $1 > \delta + e(s)$, and choose $y_i = 1$ if $1 \leq \delta + e(s)$. Then if $1 > \delta + e(s)$, everyone chooses $y_i = 0$, then the game ends

at the pro-Party 0 equilibrium $\{\hat{\alpha}^*, \hat{\beta}^*\} = \{1, 0\}$ and the new proportion of party support is $\{\alpha', \beta'\} = \{\alpha, \beta + s(1-\alpha-\beta)\}$.

If $1 \leq \delta + e(s)$, then Type 1 citizens, angered by the repression, now choose $y_i = 1$. But notice that Type 0.5 agents who do not receive the informative signal with probability $1 - s$ do not switch to Party 1 and update their δ. They still choose $y_i = 0$ given $\hat{\alpha} = 1 > 0 + \delta = \hat{\beta} + \delta$. Then the turnout and the updated expectations become $\{\hat{\alpha}', \hat{\beta}'\} = \{\alpha + (1-s)(1-\alpha-\beta), \beta + s(1-\alpha-\beta)\}$. This corresponds to a "civil war" under which society is divided along two polarized political camps with no moderate voice, even though $(1-s)(1-\alpha-\beta)$ of them are closet moderates.

We analyse whether this civil war could be an equilibrium. Notice first that since $1 \leq \delta + e(s)$ and $\hat{\alpha}' < 1$, then $\hat{\alpha}' < \delta + e(s) + \hat{\beta}'$. Type 1 citizen continues to choose $y_i = 1$. Type 0.5 citizen chooses $y_i = 0$ if $\hat{\alpha}' > \hat{\beta}' + \delta$, or $\alpha + (1-s)(1-\alpha-\beta) > \beta + s(1-\alpha-\beta) + \delta$. Rearrange and we obtain the condition $s < \frac{\alpha-\beta-\delta}{2(1-\alpha-\beta)} + \frac{1}{2} = \frac{1-2\beta-\delta}{2(1-\alpha-\beta)}$. This suggests that the repression/shock $r = s$ must be small enough, and the range of repression that satisfies this condition is decreasing in β, increasing in α, and decreasing in δ. In particular, it is impossible for Type 0.5 to choose $y_i = 0$ if $1-2\beta-\delta \leq 0$ or $\beta \geq \frac{1-\delta}{2}$. If the actual proportion of Party 1 supporters and integrity are large enough, Type 0.5 will not stick to the regime.

The condition for Type 0.5 citizens to choose $y_i = 0.5$ is $\left|\hat{\alpha}' - \hat{\beta}'\right| \leq \delta$, or $-\delta \leq \alpha - \beta + (1-2s)(1-\alpha-\beta) \leq \delta$. Rearrange and we find $\frac{1-2\beta+\delta}{2(1-\alpha-\beta)} \geq s \geq \frac{1-2\beta-\delta}{2(1-\alpha-\beta)}$. The condition is therefore $\min\{1, \max\{0, \frac{1-2\beta+\delta}{2(1-\alpha-\beta)}\}\} \geq s \geq \max\{0, \frac{1-2\beta-\delta}{2(1-\alpha-\beta)}\}$. Obviously the range is increasing in integrity δ. The upper and lower bounds are both increasing in α if $\beta < \frac{1-\delta}{2}$. A ceteris paribus increase in α will increase the upper bound more than the lower bound, increasing the range ($\frac{\partial}{\partial\alpha}\frac{1-2\beta+\delta}{2(1-\alpha-\beta)} > \frac{\partial}{\partial\alpha}\frac{1-2\beta-\delta}{2(1-\alpha-\beta)}$) unless the upper bound is 1 or $\frac{1-2\beta+\delta}{2(1-\alpha-\beta)} \geq 1$; rearrange and we get $\alpha \geq \frac{1-\delta}{2}$. If β is small, the range is increasing with α until it reaches $\frac{1-\delta}{2}$ and begins to drop. If $\frac{1-\delta}{2} < \beta < \frac{1+\delta}{2}$, the upper bound is increasing in α until it reaches 1 (when $\alpha \geq \frac{1-\delta}{2}$) while the lower bound remains zero. If $\beta > \frac{1+\delta}{2}$, then the upper bound is zero and Type 0.5 citizens will not choose $y_i = 0.5$. This result is intuitive in that as long as α, the true strength of Party 0, is small and β is not too large, the possible range in which the moderates will speak out their true preference is the largest. If either of α and β gets large, the moderates' room will be squeezed and they will falsify their preferences by supporting the camp that seems more popular.

The condition for Type 0.5 to choose $y_i = 1$ is $s > \frac{1-2\beta+\delta}{2(1-\alpha-\beta)}$, which is possible only if $\frac{1-2\beta+\delta}{2(1-\alpha-\beta)} < 1$ or $\alpha < \frac{1-\delta}{2}$. As discussed if $\beta > \frac{1+\delta}{2}$, they will choose $y_i = 1$ given any shock. The lower bound is increasing and therefore the range is decreasing in δ and α. Type 0 citizens choose $y_i = 1$ if $\hat{\alpha}' < \hat{\beta}' - \delta$, or $\alpha + (1-s)(1-\alpha-\beta) < \beta + s(1-\alpha-\beta) - \delta$, or $s > \frac{1-2\beta+\delta}{2(1-\alpha-\beta)}$, the same as the condition for Type 0.5. Otherwise they will continue stand by the regime.

Now we find the revised equilibrium $\{\hat{\alpha}'', \hat{\beta}''\}$. The civil war is only an equilibrium if Type 0.5 chooses $y_i = 0$, or when $s < \frac{1-2\beta-\delta}{2(1-\alpha-\beta)}$ and $\beta < \frac{1-\delta}{2}$. In this case, because when the opposition has a weak base (compared to δ) and the shock is too small, Party 0 is able to solidify its support at $\alpha+(1-s)(1-\alpha-\beta)$ by recruiting moderates who did not receive the negative signal from repression. The equilibrium is $\{\alpha+(1-s)(1-\alpha-\beta), \beta+s(1-\alpha-\beta)\}$. This new type of equilibrium is created by the divergence of δ between Party 1 supporters and that of the moderates due to the emotional effect $e(r)$.

Otherwise, if $\frac{1-2\beta+\delta}{2(1-\alpha-\beta)} \geq s \geq \frac{1-2\beta-\delta}{2(1-\alpha-\beta)}$, Type 0.5 will choose $y_i = 0.5$, Type 1 will choose $y_i = 1$, and Type 0 will choose $y_i = 0$. $\{\hat{\alpha}'', \hat{\beta}''\}$= $\{\alpha, \beta + s(1-\alpha-\beta)\}$. Then we will end at the truthful equilibrium if $\left|\hat{\alpha}'' - \hat{\beta}''\right| = \left|\alpha - \beta - s(1-\alpha-\beta)\right| \leq \delta$. If not, then since we know $\hat{\alpha}' > \hat{\alpha}''$, $\hat{\beta}' = \hat{\beta}''$ and $\left|\hat{\alpha}' - \hat{\beta}'\right| \leq \delta$, $\hat{\alpha}'' < \hat{\beta}'' - \delta$, then the moderates will choose $y_i = 1$ and Party 0 supporters $y_i = 1$, resulting in the Pro-Party 1 equilibrium. The regime collapses. If $s > \frac{1-2\beta+\delta}{2(1-\alpha-\beta)}$ and $\alpha < \frac{1-\delta}{2}$, everyone chooses $y_i = 1$ and we end at the pro-Party 1 equilibrium.

The equilibrium conditions are summarized as follows:

$$\{\alpha^*, \beta^*\} = \begin{cases} \{1, 0\} & \text{if } \delta + e(s) < 1 \\ \{\alpha + (1-s)(1-\alpha-\beta), \beta + s(1-\alpha-\beta)\} & \text{if } \delta + e(s) \geq 1, s < \frac{1-2\beta-\delta}{2(1-\alpha-\beta)} \\ \{\alpha, \beta + s(1-\alpha-\beta)\} & \text{if } \delta + e(s) \geq 1, \frac{1-2\beta+\delta}{2(1-\alpha-\beta)} \geq s \geq \frac{1-2\beta-\delta}{2(1-\alpha-\beta)} \\ & \left|\alpha - \beta - s(1-\alpha-\beta)\right| \leq \delta \\ \{0, 1\} & \text{if } \delta + e(s) \geq 1, \frac{1-2\beta+\delta}{2(1-\alpha-\beta)} \geq s \geq \frac{1-2\beta-\delta}{2(1-\alpha-\beta)}, \\ & \alpha < \beta + s(1-\alpha-\beta) - \delta \\ \{0, 1\} & \text{if } \delta + e(s) \geq 1, s > \frac{1-2\beta+\delta}{2(1-\alpha-\beta)} \end{cases}$$

It is apparent that once the regime "overshoots" its repression by choosing r such that $\delta + e(r) \geq 1$, it will generally lose public support by inducing angered Type 1 citizens to openly oppose the regime and some 0.5 types to become Type 1 citizens. If the repression is large enough and Party 1 has substantial support, the spiralling shock may induce other moderates and Party 0 supporters to jump on the revolutionary bandwagon, resulting in the downfall of the regime. This implies that even though dictators can employ force at will, they should only use it when the shock is actually equilibrium-changing and if using force yields a better outcome than otherwise.

4.2 Dictator with Perfect Information

Now we consider a dictator who perfectly understands the consequences of repression and knows α, β, δ, $e(r)$, and $p(r) = r$. I assume that when repression and no repression yields the same outcome in $\hat{\alpha}$, the dictator prefers to not repress to avoid increasing β. This

is reasonable since under a dynamic setting, a dictator anticipating future shocks would prefer not to alienate the moderates and strengthen Party 1 by unnecessarily repressing small protests now. The immediate implication is that if $s < \frac{1-\delta}{2}$, Party 0 will not repress because the shock is only transient: Type 1 citizens will not change their public preference y_i upon observing $\{1-s, s\}$. Then a dictator with perfect information will do better than a naive dictator if there is some range of $s < \frac{1-\delta}{2}$ such that $\delta + e(s) \geq 1$.

We now consider two separate cases: if $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, which under no repression will switch the equilibrium to $\{\alpha, \beta\}$, and if $s > \frac{1+\delta}{2}$, which will result in a $\{0, 1\}$ equilibrium if Party 0 does not repress. In these two scenarios the dictator generally will not fully repress the shock. Under a moderate shock the dictator may not repress and voluntarily allows the truthful equilibrium. Under a high shock, however, the dictator always chooses some level of repression.

4.2.1 Moderate shock: $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$

Facing moderate shock $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, Party 0 compares the outcomes over the range of possible repression $r \in [0, \min\{s, R\}]$ before choosing optimal repression. For any repression r, Type 1 citizens choose $y_i = 0$ under $\{\hat{\alpha}, \hat{\beta}\} = \{1-s+r, s-r\}$ and $e = e(r)$ if $U(y_i = 0|x_i = 1) - U(y_i = 1|x_i = 1) > 0$, or $1 - s + r - (\delta + e(r)) > s - r$. Rearranging we obtain $r > s - \frac{1-\delta}{2} + \frac{e(r)}{2}$. I define $s - \frac{1-\delta}{2} + \frac{e(r)}{2}$ to be the *threshold* of optimal repression. Any r greater than this threshold will induce every citizen to choose $y_i = 0$.

Note that since $r \leq s$, that repression is possible requires $e(r) < 1 - \delta$. This is obvious since that the integrity parameter $\delta + e(r) \geq 1$ means that the individual will never engage in preference falsification. Furthermore, it implies that optimal repression is always less than s if feasible. A smart dictator will not exert maximum repression and leaves some breathing room for the opposition, which will vanish if it does not gain enough sympathy from being repressed. An increase in δ will reduce the range where $e(r) < 1 - \delta$ holds since e is increasing in r. High integrity makes Type 1 citizens express $y_i = 1$ under any repression.

Proposition 1 *Under moderate shock $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, optimal repression, defined as the repression needed to restore equilibrium back to $\{1, 0\}$, exists if and only if there exists $r \in [0, \min\{s, R\}]$ such that $e(r) < 1 - \delta$ and $r > s - \frac{1-\delta}{2} + \frac{e(r)}{2}$.*

The inequality $r > s - \frac{1-\delta}{2} + \frac{e(r)}{2}$ provides another implicit restriction on the $e(r)$ function: for every $k \in [0, \delta]$ there must be some $e(r) < 2(r-k)$ on $[0, \min\{s, R\}]$ for optimal repression to exist for $s = \frac{1-\delta}{2} + k$. These conditions give us some necessary and sufficient conditions for the shape of $e(r)$ under given δ for optimal repression to exist.

Proposition 2 *Assume that $e(r)$ is differentiable on $[0, \min\{\frac{1+\delta}{2}, R\}]$. Suppose there is a $\epsilon > 0$*

such that for all $r \in [0, \delta + \epsilon]$, $e'(r) < \frac{2\epsilon}{\delta+\epsilon}$. *Then optimal repression exists for all* $s \in [\frac{1-\delta}{2}, \frac{1+\delta}{2}]$.

Proof. Consider $r = \delta + \epsilon$ and $s = \frac{1+\delta}{2}$. By the Mean Value Theorem, $e(\delta + \epsilon) - e(0) = e'(c)(\delta + \epsilon)$ for some c in $[0, \delta + \epsilon]$. $e(0) = 0$ and $e'(c) < \frac{2\epsilon}{\delta+\epsilon}$ implies $e(r) < 2(r - \delta)$ which satisfies the optimal repression threshold. In fact, the Mean Value Theorem implies that for any $k \leq \delta$, $e(k+\epsilon) - e(0) < \frac{2\epsilon}{\delta+\epsilon}(k+\epsilon) < 2\epsilon = 2(k+\epsilon-k)$. Then the optimal repression for $s = k + \frac{1-\delta}{2}$ lies in $(k, k + \epsilon]$ for all $k \in [0, \delta]$ since a dictator with perfect information wants to minimize also the information effect on Type 0.5 citizens.

Intuitively, this sufficient condition implies that $e(r)$ must not increase too fast with r and for optimal repression to exist for larger values of s, $e(r)$ generally has to increase more slowly. If Type 1 citizens are not easily angered by repression, then the fear effect dominates the emotional effect. Note that this condition is not necessary: we can imagine $e(r)$ that is increasing slowly near the origin before increasing rapidly near optimal repression $r_k = k + \epsilon$ for shock $s = \frac{1-\delta}{2} + k$, yet $e(r_k) < 2(r_k - k)$. Alternatively, $e(r)$ could be increasing rapidly at the origin and flattening out at r_k before hitting $2(r_k - k)$. The key requirement is that $e(r) < 2(r - k)$ for some $s = \frac{1-\delta}{2} + k$. If Party 0 implements optimal repression, all citizens choose $y_i = 0$, $\{\hat{\alpha}', \hat{\beta}'\} = \{1, 0\}$, $\{\alpha, \beta\} = \{\alpha, \beta + r(1 - \alpha - \beta)\}$. The crisis is resolved at optimal repression r_k for $s = k + \frac{1-\delta}{2}$, and the game ends at the pro-Party 0 equilibrium.

To further illustrate, consider Figure 1. Figure 1 is a composition of two graphs: the graph on the top (with s, r as the x-axis) measures the strength of two shocks s, s' and corresponding repressions r, r'. The direction of the arrows indicate whether they represent a shock (pointing towards the right) or a repression (pointing towards the left). Shock s, s' are both in $[\frac{1-\delta}{2}, \frac{1+\delta}{2}]$, with $s' > s$.

Repression level r is a candidate for optimal repression in response to shock s since $s - r < \frac{1-\delta}{2}$, which always makes Type 0.5 citizens choose $y_i = 0$. If $s - r \geq \frac{1-\delta}{2}$, then Type 0.5 citizens will choose $y_i = 0.5$ and hence Type 1 citizens will choose $y_i = 1$, in which case the repression is not effective. r can be decomposed into parts k and ϵ, where k represents the amount needed for $s - r = \frac{1-\delta}{2}$ and ϵ is the marginal repression required for $s - r < \frac{1-\delta}{2}$. Similarly, repression level r' is a candidate for optimal repression in response to shock s'.

Whether r, r' are optimal repressions depends on the shape of $e(r)$. Consider the graph in the bottom which y-axis is $e(r)$ and x-axis is r. Four different $e(r)$ functions, $e_1(r), e_2(r), e_3(r), e_4(r)$ are drawn in the graph. Mapping r, r', ϵ from the s, r graph to the $r - e(r)$ graph, we see that r satisfies the threshold condition for s under e_1, e_2, e_3, e_4 since $e_i(r) < 2(r - k) = 2\epsilon$ for all $i = 1, 2, 3, 4$. e_2, e_4 are linear and hence they definitely satisfy the assumption of Proposition 2 ($e'(r) < \frac{2\epsilon}{\delta+\epsilon}$). e_1 is convex and e_3 is concave, yet under $e = e_1$ or e_3, r remains an optimal repression for s.

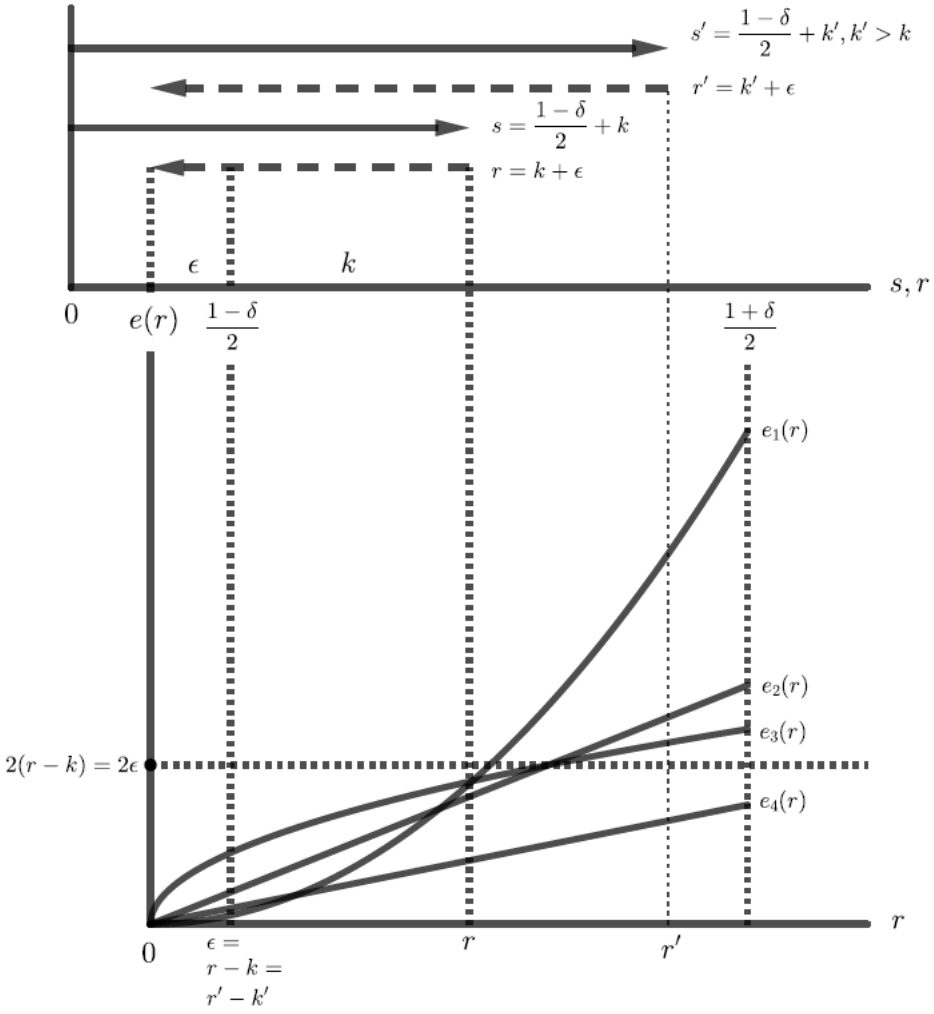

Figure 1: Mapping repression to e(r), moderate shock

However r' is not an optimal repression for s' under e_1, e_2, e_3 since $e_i(r') > 2\epsilon$, $i = 1, 2, 3$. Under e_4, which has a flatter slope than e_2, $e_4(r') < 2\epsilon$ and hence r' is an optimal repression level for s'. This demonstrates the intuition that as the shock gets larger and the corresponding repression increases, $e(r)$ has to increase slowly for optimal repression to exist.

If optimal repression does not exist for s, then Party 0 considers whether a repression level that makes Type 0.5 citizens support the Party is still available, which I term as *second best repression*. Making Type 0.5 choose $y_i = 0$ requires $U(y_i = 0|x_i = 0.5) - U(y_i = 0.5|x_i = 0.5) > 0$, or $1 - s + r - 0.5\delta > 0.5(1 - s + r) + 0.5(s - r)$, $r > s - \frac{1-\delta}{2}$. $s - \frac{1-\delta}{2}$ is the threshold for effective second best repression.

Since Party 0 recognizes the information effect, it would want to reduce r as small as possible and for shock s it will choose $r = s - \frac{1-\delta}{2} + \epsilon$, where $\epsilon > 0$ can be interpreted

as the smallest non-divisible unit of repression. Returning to Figure 1, we see that r, r' are automatically second best repressions for shocks s, s'. Again the second best repression will always be less than the shock. Then the updated expectation after repression is $\{\hat{\alpha}', \hat{\beta}'\} = \{\alpha + (1 - (s - \frac{1-\delta}{2} + \epsilon))(1 - \alpha - \beta), \beta + (s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta)\}$.

Under $\{\hat{\alpha}', \hat{\beta}'\}$, Type 1 citizen chooses $y_i = 0$ if $\hat{\alpha}' - \delta - e(s - \frac{1-\delta}{2} + \epsilon) > \hat{\beta}'$. Plugging in the expressions and collecting terms we get

$$\delta + e(s - \frac{1-\delta}{2} + \epsilon) < 1 - 2\beta - 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) \tag{5}$$

Observe first that since we have assumed optimal repression does not exist for s, there is no repression level that satisfies both $e(r) < 1 - \delta$ and $r > s - \frac{1-\delta}{2} + \frac{e(r)}{2}$. Then, under second best repression $r = s - \frac{1-\delta}{2} + \epsilon$, $e(r) = e(s - \frac{1-\delta}{2} + \epsilon) \geq 2\epsilon$, yet (5) holds. Therefore,

$$2\epsilon \leq e(s - \frac{1-\delta}{2} + \epsilon) < 1 - 2\beta - 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) - \delta \tag{6}$$

This situation cannot hold if $2\beta + 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) + \delta \geq 1$ since the right-hand side will not be positive and hence will not be greater than 2ϵ. The higher the shock s, integrity δ, the true support of the opposition β and that of the moderates $1 - \alpha - \beta$, the less likely the condition is to hold. If it does hold, "second best repression" eventually will induce Party 1 supporters to support Party 0. Since the criterion for Type 1 to choose $y_i = 0$ is more stringent than Type 0.5, it follows that Type 0.5 will also choose $y_i = 0$. Intuitively, this corresponds to a situation where the repression makes Type 1 supporters angry enough to protest on the streets initially, but their strength turn out to be insufficient (low $\beta + (s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta)$). If integrity δ is low enough so that (6) holds, in the next period Type 1 supporters will return home and pretend to support the regime again. It will appear that the people have "forgotten" about the repression, but in reality they remember but choose to remain silent. Then Party 0 will simply implement this repression level $r = s + \frac{\delta-1}{2} + \epsilon$ and get the optimal repression outcome.

Furthermore, notice that this is the *minimum* possible level of optimal repression since Type 0.5 citizens will no longer support Party 0 if r is set slightly lower and Type 1's threshold for $y_i = 0$ is always at least as high as Type 0.5's. Then this implies that a very flat $e(r)$ will make repression easy for the dictator. Returning to Figure 1, we can see that for $e_4(r)$, the e function that increases the slowest, optimal repression exists for all $s \in [\frac{1-\delta}{2}, \frac{1+\delta}{2}]$. The existence of optimal repression again hinges critically on whether $e(r)$ increases relatively flatly. The degree to which repression angers and mobilizes sympathizers of the opposition matters in whether the dictator could be effectively constrained.

Proposition 3. *Suppose now that for shock $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, $e(s - \frac{1-\delta}{2} + \epsilon) < 1 - 2\beta - 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) - \delta$ for some $\epsilon > 0$. Then optimal repression exists and equals $s - \frac{1-\delta}{2} + \epsilon$. This is the minimum level of optimal repression.*

If this condition fails, Type 1 citizens will continue choosing $y_i = 1$ after observing $\{\hat{\alpha}', \hat{\beta}'\}$. For Type 0.5 citizens, they will choose $y_i = 0$ if and only if $\hat{\alpha}' - 0.5\delta > 0.5(\hat{\alpha}' + \hat{\beta}')$, or $\alpha + (1 - (s - \frac{1-\delta}{2} + \epsilon))(1 - \alpha - \beta) - 0.5\delta > 0.5$, rearrange and we get the condition that

$$\delta < 1 - 2\beta - 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) \tag{7}$$

which is the condition for Type 1 in (5) absent the $e(r)$ term.

If Type 0.5 citizens further revise their expectations and abandon Party 0 following second-best repression, the result is no better and could be worse than the truthful equilibrium for Party 0 since some proportion of the moderates has now switched to Party 1. Then unless (7) holds, by backward induction Party 0 prefers the truthful equilibrium and will not repress at all. This is analogous to a situation where a dictatorship in face of rising demands voluntarily democratizes. The incentive for the dictator in this model differs from Acemoglu and Robinson (2006) which emphasizes economic costs of a revolution and the threat of redistribution from below. Here the regime anticipates that repression will be futile since it knows it is highly unpopular, the people have high integrity, and they are no longer fearful due to the powerful signal the shock has sent about the opposition's strength. In reaction, Party 0 chooses to share power with the opposition rather than risk losing further support by suppressing dissent. This explains why some popular and widespread protests such as the U.S. Civil Rights Movement in the 1960s or the peaceful Eastern European revolutions in 1989 induce governments to accept reforms or relinquish control voluntarily without much violent struggle.

Furthermore, by comparing (5) and (7) we immediately see that a civil war equilibrium holds only if $e(s - \frac{1-\delta}{2} + \epsilon)$ drives a wedge between the preferences of Type 1 and Type 0.5 citizens such that (5) fails but (7) is satisfied. The slope at which $e(r)$ increases therefore matters: larger differences between how the supporters of the opposition and the moderates perceive the regime's repressive actions will lead to polarization. The wedge may encourage the regime to repress if the strength of the opposition is revealed to be insufficient for the moderates to abandon supporting Party 0 (low β). This covers situations under which dictators like Syria's Assad choose to resist an apparently strong revolutionary movement and risk a protracted civil war.

Proposition 4. *Under moderate shock $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, if there is no optimal repression, (7) holds but (6) fails, then the dictator chooses repression $s + \frac{\delta-1}{2} + \epsilon$ to force a civil war equilibrium where $\{\hat{\alpha}, \hat{\beta}\} = \{\alpha + (1 - s - \frac{\delta-1}{2} - \epsilon)(1 - \alpha - \beta), \beta + (s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta)\}$.*

Proposition 5. *Under moderate shock $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, if there is no optimal repression, (7) fails, then the dictator chooses repression $r = 0$ to end the game in the truthful equilibrium $\{\alpha, \beta\}$.*

4.2.2 High shock: $s > \frac{1+\delta}{2}$

Under high shock, Party 0 is willing to repress so long as the outcome is better than $\{0, 1\}$. This creates the paradoxical result that dictatorships may be more willing to repress if the shock is high while under moderate shocks the dictator may not repress. Or, in more extreme form, it *never* voluntarily democratizes if the shock left unchecked will result in an alternative autocracy. It is sensible because Party 0 does not want to be eliminated by the high shock, which will make its supporters jump ship and give Party 1 dictatorial powers. On the other hand, under moderate shock sharing power is feasible and may be the best alternative. An example of a rational dictatorship choosing to share power and keep some political influence is Myanmar's military which democratizes in exchange for representation in the legislature and the cabinet.

For Type 1 citizens to choose $y_i = 0$ after repression, we still have the threshold condition that $r > s + \frac{\delta-1}{2} + \frac{e(r)}{2}$. Under high shock $s = \frac{1+\delta}{2} + k$ for $k \in (0, \frac{1-\delta}{2}]$, this implies $r > \delta + k + \frac{e(r)}{2}$, $e(r) < 2(r - \delta - k)$. To satisfy the condition for optimal repression $e(r)$ must be increasing very slowly with r.

Proposition 6. *Under high shock $s > \frac{1+\delta}{2}$, optimal repression exists only if for shock $s = \frac{1+\delta}{2} + k$, $k \in (0, \frac{1-\delta}{2}]$ and there exists $r \in [0, \min\{s, R\}]$ such that $e(r) < 1 - \delta$ and $r > \delta + k + \frac{e(r)}{2}$. Furthermore, suppose that $e(r)$ is differentiable, $R > \frac{1+\delta}{2}$, and there exists an $\epsilon > 0$ such that for all $r \in [0, \frac{1+\delta}{2} + \epsilon)$, $e'(r) < \frac{2\epsilon}{\frac{1+\delta}{2}+\epsilon}$. Then optimal repression exists for all $s \in [0, 1]$.*

Proof. Consider the largest possible shock $s = 1$ and suppose the dictator can choose $r = \frac{1+\delta}{2} + \epsilon \leq R$. $e(r) - e(0) = e'(c)(\frac{1+\delta}{2} + \epsilon)$ for some $c \in (0, r)$ by the Mean Value Theorem. Applying $e'(c) < \frac{2\epsilon}{\frac{1+\delta}{2}+\epsilon}$ we obtain $e(r) < 2\epsilon = 2(r - \frac{1+\delta}{2})$, which satisfies the threshold condition.

The proposition shows that for optimal repression to exist for a high shock level, not only $e(r)$ must be rising really slowly with r, but the maximum coercive capacity must be large enough to support high repression. Otherwise integrity δ must be low enough to reduce the repression needed.

Similarly, the condition for second best repression is $r > s - \frac{1-\delta}{2}$. After applying the second best repression $r = s - \frac{1-\delta}{2} + \epsilon$, Type 1 citizen chooses $y_i = 1$, Type 0.5 and 0 citizens choose $y_i = 0$, generating the expectations $\{\hat{\alpha}', \hat{\beta}'\} = \{\alpha + (1 - s + \frac{1-\delta}{2} - \epsilon)(1 - \alpha - \beta), \beta + (s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta)\}$.

Type 1 citizen will switch back to $y_i = 0$ if $\hat{\alpha}' - \delta > \hat{\beta}'$, $\alpha + (1 - s - \frac{\delta-1}{2} - \epsilon)(1 - \alpha - \beta) - \delta - e(s + \frac{\delta-1}{2} + \epsilon) > \beta + (s + \frac{\delta-1}{2} + \epsilon)(1 - \alpha - \beta)$, and therefore

$$e(s - \frac{1-\delta}{2} + \epsilon) < 1 - 2\beta - 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) - \delta \tag{8}$$

Comparing (8) to (5), it is clear that since second best repression under moderate shock must be less than that under higher shock, (8) is harder to satisfy. This again is the usual condition that $e(r)$ is flat, β and δ are small, and α large. If (8) holds, minimum optimal repression exists for some high shock, then there is minimum optimal repression for all medium shocks.

Type 0.5 citizens choose 1 under $\{\hat{\alpha}', \hat{\beta}'\}$ only if $\hat{\alpha}' - 0.5\delta > 0.5(\hat{\alpha}' + \hat{\beta}')$, or $\alpha + (1 - s - \frac{\delta-1}{2} - \epsilon)(1 - \alpha - \beta) - 0.5\delta > 0.5$. After some rearranging we obtain the condition

$$\delta < 1 - 2\beta - 2(s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta) \tag{9}$$

Comparing (9) to (7) by the same argument, the condition is harder to satisfy: Type 0.5 citizens are more likely after a high shock to reveal their true preference. This is especially so if β is large (surely if $\beta > \frac{1}{2}$), shock s is large, and/or δ is small. If (9) holds but (8) fails, Party 0 implements second best repression and the game ends at a civil war equilibrium.

If second best repression does not work, it is impossible to generate outcome better than the truthful equilibrium. Since Party 1 will obtain support of all citizens if there is no repression, Party 0 will repress to some degree to win back the support of Type 0 citizens. In particular, it chooses $r \geq s - \frac{1+\delta}{2}$ so that $\hat{\alpha} + \delta = 1 - s + r + \delta \geq s - r = \hat{\beta}$, Type 0 citizens choose $y_i = 0$. Under this repression Type 0.5 citizens will choose $y_i = 0.5$ since $\left|\hat{\alpha} - \hat{\beta}\right| \leq \delta$. Then $\{\hat{\alpha}', \hat{\beta}'\} = \{\alpha, \beta + (s - \frac{1+\delta}{2})(1 - \alpha - \beta)\}$, the game ends in a truthful equilibrium. The only exception is if such repression which generates expectation $\{\alpha', \beta'\} = \{\alpha, \beta + (s - \frac{1+\delta}{2})(1 - \alpha - \beta)\}$, $\left|\alpha' - \beta'\right| > \delta$ so that the truthful equilibrium is no longer feasible. Since we have assumed $|\alpha - \beta| \leq \delta$, this is only possible if the new converts to Party 1 after repression, $(s - \frac{1+\delta}{2})(1 - \alpha - \beta)$ is large enough. Conditions under which this occurs is if the shock is high, δ low enough, and the original proportion of moderates $1 - \alpha - \beta$ is high enough. In such a setting Party 0 has no hope since it has exhausted all possibilities: the shock is high, and it is too unpopular to reverse the odds by repression.

So long as $(s - \frac{1+\delta}{2})(1 - \alpha - \beta) \leq \delta$, Party 0 generally will almost always implement repression since the outcome will be better than $\{0, 1\}$. I call this level of repression *almost sure repression*. Under this situation higher integrity δ is to Party 0's advantage. As long as its supporters are loyal enough, Party 0 could survive the crisis by first demonstrating it still has political support through almost sure repression, before turning to the negotiating tables to share power with Party 1.

Figure 2 illustrates the arguments. Here, $s = \frac{1+\delta}{2} + k$ is a high shock. The dictator has two options: choosing second best repression $r = k + \delta + \epsilon$ (which could be optimal repression if $e(r)$ is low enough) or almost sure repression $r' = k$. Consider first the second best repression r and the $e(r) - r$ graph. Here $e_1(r), e_2(r), e_3(r)$ are all greater than $2(r - \delta - k)$. Then the second best repression cannot be an optimal repression under Propo-

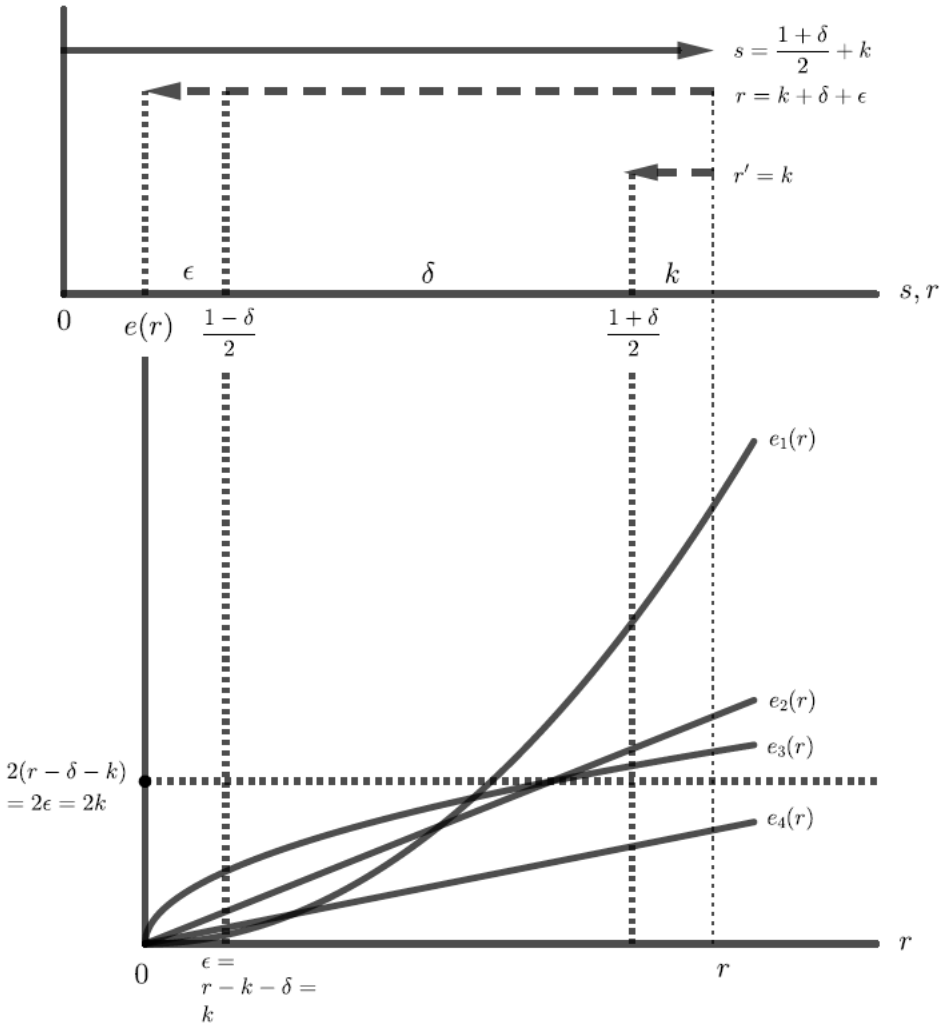

Figure 2: Mapping repression to e(r), high shock

sition 6. However, if $e = e_4$, $e_4(r) < 2(r - \delta - k)$, then under e_4 optimal repression exists even for high shock level s. This again demonstrates the critical importance of $e(r)$ in the dictator's decision to repress: if the people are apathetic enough to political violence, the regime can suppress crises at will and is constrained only by the coercive capabilities it could command.

However, if (9) does not hold, implementing r will not generate a better outcome than the truthful equilibrium. The dictator considers the almost sure repression r', which will convert less moderates to Party 1 and result in a better outcome than r. Since $r' = k = s - \frac{1+\delta}{2}) < \delta$, truthful equilibrium remains feasible after r', and the dictator will implement r' as a last resort.

To summarize, under perfect information Party 0's strategy is:

1. implement $r = 0$ if $s < \frac{1-\delta}{2}$, the game ends at pro-Party 0 equilibrium $\{1, 0\}$;

2. if $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, (6) holds or $e(s+\frac{\delta-1}{2}+\epsilon) < 2\epsilon$, choose minimum optimal repression $s + \frac{\delta-1}{2} + \epsilon$, the game ends at pro-Party 0 equilibrium $\{1, 0\}$;

3. If $s > \frac{1+\delta}{2}$, (8) holds or $e(s + \frac{\delta-1}{2} + \epsilon) < 2\epsilon$, choose minimum optimal repression $r = s + \frac{\delta-1}{2} + \epsilon$, $k = s - \frac{1+\delta}{2}$, the game ends at pro-Party 0 equilibrium $\{1, 0\}$;

4. If $s \geq \frac{1-\delta}{2}$, if there exists $r \in [0, \min\{s, R\}]$ such that $e(r) < 2(r - s + \frac{1-\delta}{2})$, implement optimal repression r to end at pro-Party 0 equilibrium $\{1, 0\}$;

5. If $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, if there is no optimal repression, (7) holds but (5) fails, choose second best repression $r = s + \frac{\delta-1}{2} + \epsilon$ to end at civil war equilibrium $\{\alpha + (1 - s + \frac{1-\delta}{2} - \epsilon)(1 - \alpha - \beta), \beta + (s - \frac{1-\delta}{2} + \epsilon)(1 - \alpha - \beta)\}$;

6. If $\frac{1+\delta}{2} \geq s \geq \frac{1-\delta}{2}$, if there is no optimal repression, (7) fails, choose $r = 0$ and the game ends at truthful equilibrium $\{\alpha, \beta\}$;

7. If $s > \frac{1+\delta}{2}$, (9) holds but (8) fails, choose second best repression $s + \frac{\delta-1}{2} + \epsilon$, the game ends at civil war equilibrium $\{\alpha+(1-s+\frac{1-\delta}{2}-\epsilon)(1-\alpha-\beta), \beta+(s-\frac{1-\delta}{2}+\epsilon)(1-\alpha-\beta)\}$;

8. If $s > \frac{1+\delta}{2}$, optimal repression is not possible, (9) fails, choose almost sure repression $s - \frac{1+\delta}{2}$, the game ends at the truthful equilibrium $\{\alpha, \beta + (s - \frac{1+\delta}{2})(1 - \alpha - \beta)\}$ unless $(s - \frac{1+\delta}{2})(1 - \alpha - \beta) \leq \delta$, under which the game ends at pro-Party 1 equilibrium $\{0, 1\}$.

Shock level	small $e(r)$	large $e(r)$, high β	large $e(r)$, low β, low δ
low	not repress, keep power		
medium	repress, keep power	not repress, share power	repress, civil war
high	repress, keep power	repress, share power	repress, civil war

Table 1: The dictator's strategy in words

Compared to the naive dictator case, we see that having information about the actual proportions of support and the citizens' utility parameters produces better outcomes for the incumbent regime under almost all scenarios. By repressing *just enough* rather than executing a complete crackdown, the dictator avoids unnecessarily provoking sympathizers of the opposition while creating enough impression of strength to deter citizens from protesting. When the shock and the opposition's real strength is too strong, the regime also has the opportunity to make democratic compromises that prevent the worst outcome of complete annihilation by the opposition party. In reality, dictatorships generally do not have perfect information about the relevant parameters but they recognize that repressions can be counterproductive and that people falsify their preferences. In the next section I discuss the implications of regimes acting on the strategy delineated above without complete information about $\alpha, \beta, \delta, e(r)$, and $p(r)$.

4.3 Dictator with Imperfect Information

Finally, I consider a Party 0 that has "optimistic" estimates of $\bar{\alpha} > \alpha$, $\bar{\beta} < \beta$, $\bar{\delta} < \delta$, $\bar{e}(r) < e(r)$ for all r, and $\bar{r} < r$ in a stylized manner. A rational dictatorship follows the strategy adopted by the fully informed dictator using its own estimates of the parameters. Then if $s < \frac{1-\bar{\delta}}{2}$, Party 0 will not repress. The immediate consequence is that for shock $s \in [\frac{1-\delta}{2}, \frac{1-\bar{\delta}}{2})$, the dictator may "undershoot" the optimal level of repression by choosing not to repress an apparently small shock that actually would produce big consequences: it could bring about a truthful equilibrium. This accords with Kuran (1989)'s stories of overconfident monarchs choosing to ignore gathering protests before falling prey to sudden revolutions. In contrast, a "pessimistic" or "paranoid" regime that overestimates δ will repress shocks that are actually insignificant. Under certain circumstances an $e(r)$ that increases fast enough will allow these unnecessary repressions to generate large-scale revolutions, confirming the regime's paranoia.

If $s \in [\frac{1-\bar{\delta}}{2}, \frac{1+\bar{\delta}}{2}]$, again the dictator chooses optimal repression if for $s = \frac{1-\bar{\delta}}{2} + k$ there is r such that $\bar{e}(r) < 2(r-k)$ and $\bar{e}(r) < 1 - \bar{\delta}$. Since $e(r) > \bar{e}(r)$, $\delta > \bar{\delta}$, again there is room for "mistakes" that the regime may choose some r such that $e(r) \geq 2(r-k)$, which will only lead Type 1 citizens to continue opposing the regime and more moderates joining Party 1. Similarly, if the regime opts for second best repression $r = s - \frac{1-\bar{\delta}}{2} + \epsilon$, it considers equations (5) and (7) in terms of its own estimates:

$$\bar{\delta} + \bar{e}(s - \frac{1-\bar{\delta}}{2} + \epsilon) < 1 - 2\bar{\beta} - 2\bar{p}(s - \frac{1-\bar{\delta}}{2} + \epsilon)(1 - \bar{\alpha} - \bar{\beta}) \tag{10}$$

$$\bar{\delta} < 1 - 2\bar{\beta} - 2\bar{p}(s - \frac{1-\bar{\delta}}{2} + \epsilon)(1 - \bar{\alpha} - \bar{\beta}) \tag{11}$$

The left hand side is smaller and the right hand side is larger in (10) and (11) compared to their real-world counterparts (5) and (7), which again imply that the optimistic dictator may misjudge the situation: if (10) holds but (5) fails, the dictator chooses $r = s - \frac{1-\bar{\delta}}{2} + \epsilon$, the minimal optimal repression from his perspective. However, the level is insufficient since $s - \frac{1-\bar{\delta}}{2} + \epsilon < s - \frac{1-\delta}{2} + \epsilon$, the "correct" level of minimal optimal repression. Then as a consequence of undershooting repression, Type 1 and Type 0.5 citizens will still choose to express their private preferences, bringing about a truthful equilibrium with $s - \frac{1-\bar{\delta}}{2} + \epsilon$ of the moderates joining Party 1. The story is similar if (11) holds but (7) fails: the dictator chooses second best repression that is actually insufficient to induce Type 0.5 citizens to choose $y_i = 0$. This also implies that a dictator optimistic about his own popularity, even though not completely ignorant, is less likely to voluntarily democratize when faced with a moderate shock.

Since $\frac{1+\bar{\delta}}{2} < \frac{1+\delta}{2}$, the optimistic regime with imperfect information may mistakenly consider a moderate shock to be a high shock and choose to repress for $s \in [\frac{1+\bar{\delta}}{2}, \frac{1+\delta}{2}]$ even when not repressing may be a better option. Under higher shock levels the optimistic dic-

tator is less likely to concede and democratize since from his point of view δ and $e(r)$ is lower than it actually is, suggesting that coercion is effective. In contrast, a pessimistic regime may mistake a high shock for a moderate shock and try to share power by peacefully democratizing. A pessimistic Party 0's overconfidence in its people's integrity may result in the opposition party gaining complete power.

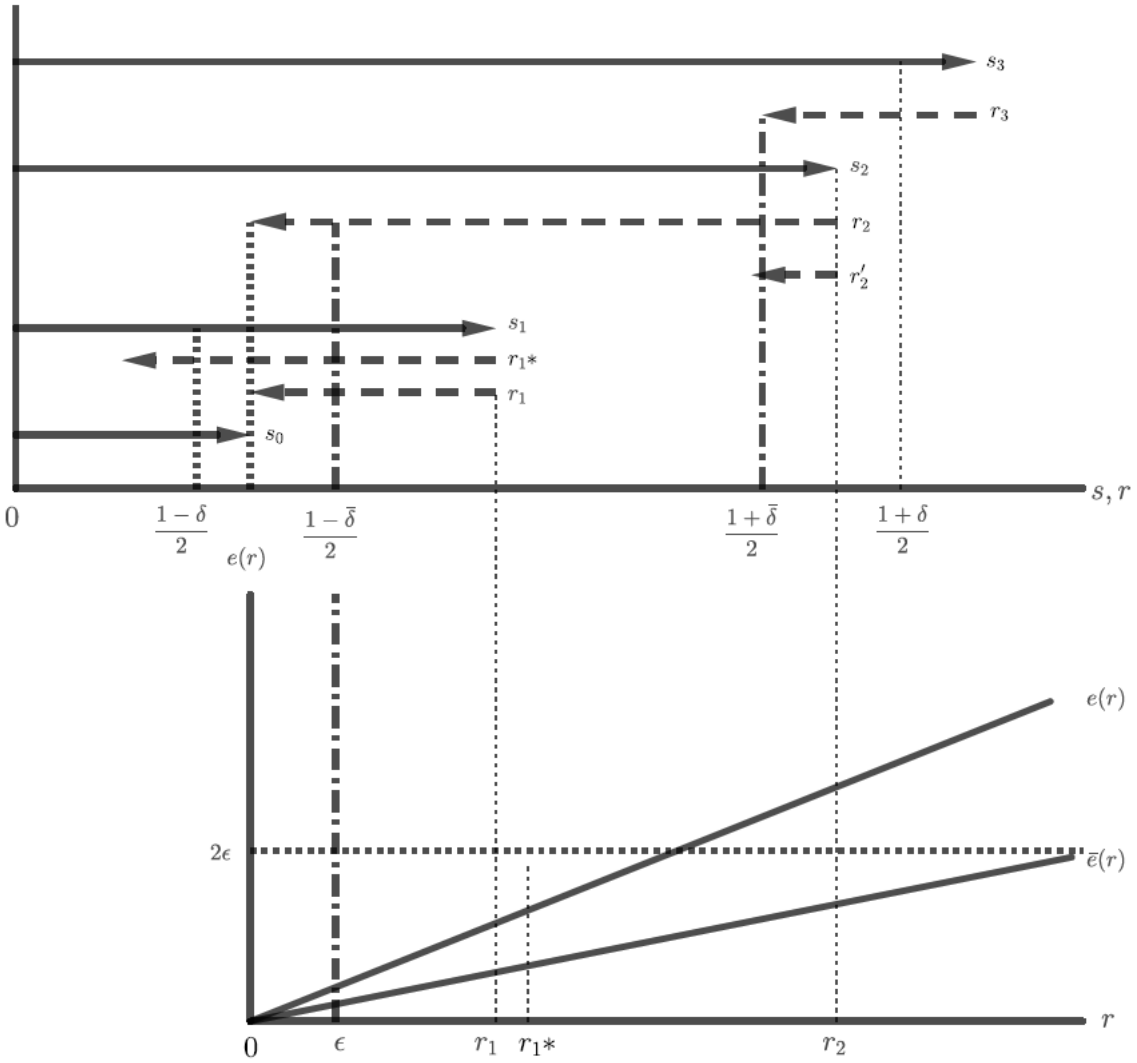

Figure 3: "Mistakes" of an optimistic dictator

Figure 3 describes the actions of an optimistic dictator that are sub-optimal. Since $\bar{\delta} < \delta$, $\frac{1-\delta}{2} < \frac{1-\bar{\delta}}{2} < \frac{1+\bar{\delta}}{2} < \frac{1+\delta}{2}$. On the $r - e(r)$ graph, $\bar{e}(r) > e(r)$ for all r. First consider shock s_0 at the bottom of the s, r graph. Since $s_0 > \frac{1-\delta}{2}$, it is a moderate shock that will induce a truthful equilibrium if not suppressed. However, since $s_0 < \frac{1-\bar{\delta}}{2}$, the dictator considers it a small shock and will choose $r = 0$. Then Type 1 and 0.5 citizens reveal their true preferences, resulting in a revolution.

Facing shock s_1, a larger moderate shock such that $s_1 > \frac{1-\bar{\delta}}{2}$, the dictator chooses

$r_1 = s_1 - \frac{1-\bar{\delta}}{2} + \epsilon$ such that $\bar{e}(r_1) < 2\epsilon$. However since $s_1 - r_1 > \frac{1-\delta}{2}$, the repression level is insufficient to persuade Type 1 and Type 0.5 supporters to choose $y_i = 0$ and will still bring about a truthful equilibrium. Furthermore, an additional proportion r_1 of the moderates turn to Party 1, making r_1 a much worse option than optimal repression $r^* = s_1 - \frac{1-\delta}{2} + \epsilon$ (optimal here since $e(r^*) < 2\epsilon$) or even no repression $r = 0$.

Now consider an even larger moderate shock s_2, which is a high shock from the dictator's point of view. The dictator chooses $r_2 = s - \frac{1-\bar{\delta}}{2} + \epsilon$, an optimal repression from his perspective since $\bar{e}(r_2) < 2(r_2 - s + \frac{1-\bar{\delta}}{2} - \epsilon)$. However, it is not an optimal repression: $r_2 < s - \frac{1-\delta}{2} + \epsilon$, so it is insufficient to convince both Type 1 and Type 0.5 citizens to choose $y_i = 0$. Furthermore, since also $e(r_2) > 2\epsilon$, the backlash from the repression will be higher than what the optimistic regime expects. Then it will end up in a double whammy: the repression fails to silence the opposition and will provoke substantial backlash. Finally, for a real high shock s_3, the regime implements almost sure repression $r_3 = s_3 - \frac{1+\bar{\delta}}{2}$ that is larger than the level actually needed ($s_3 - \frac{1+\delta}{2}$), which will push more moderates to join the opposition.

If the regime's estimates are closer to that in reality, then the probability of committing mistakes and the margins of error are generally smaller. This explains why sophisticated authoritarian regimes like the Chinese Communist Party try to understand their citizens' true preferences by investing heavily in surveillance systems and letting spies and party bureaucrats infiltrate networks of common citizens (Edmond, 2013). Better information allows dictators to make better decisions when faced with unexpected shocks. At the same time, authoritarian regimes that understand the logic of preference falsification also spend on propaganda and censorship to create false perceptions of their popularity that could be self-confirming (Guriev & Treisman, 2019).

5 Conclusion

In this paper I analyze the implications of preference falsification on the dictator's decision to repress an unexpected shock to regime stability, under the assumption that repression could yield three different effects: imposing fear and restoring apparent regime popularity, angering supporters of the opposition, and causing the political moderates to sympathize with the opposition. In contrast to influential theoretical work which explains variation in political outcomes by economic determinants (Acemoglu & Robinson, 2006; Besley & Persson, 2011), I show that the variation can also be a function of *psychological* determinants like δ, the cost of preference falsification which proxies the degree of political tolerance in the society, and $e(r)$, citizens' emotional response to violent repression.

In particular the model suggests that

1. small shocks should be tolerated;

2. optimal repression exists for moderate or high shocks if the emotional response to repression is small;

3. if the emotional response to repression is high so that repression is counterproductive, the regime voluntarily democratizes when facing moderate shocks;

4. civil war exists if repression induces a strong enough emotional effect to drive a large wedge between the moderates' and the opposition's threshold to publicly support the regime;

5. the regime almost always represses under high shock to preserve strength.

6. Since information on true popularity and citizens' decision utility is unknown, regimes routinely make mistakes in over-repressing or under-repressing based on their flawed estimates.

What generates the dictator's optimal strategy is the payoff *discontinuity* arising from the citizens' thresholds of preference falsification/revelation, which are determined by both the exogenous parameters δ, β, α, s and the endogenous choice of repression level r, $e(r)$. A small margin of error in the choice of repression produces vastly disparate outcomes, even holding all relevant factors constant. The results echo Kuran (1989)'s emphasis on the inherent unpredictability of revolutions and suggests that the source of such randomness lies in the unobserved decision rule of citizens in choosing their public preferences.

Another interesting implication is that in societies where δ is high, rational authoritarians repress more since the threshold for a shock to threaten the regime, $\frac{1-\delta}{2}$, is decreasing with δ. In low δ societies the regime may actually choose not to repress protests that are too small to persuade citizens to reveal their true preferences. On the other hand, dictators of low δ societies are less likely to yield to large shocks and are willing to repress or engage in a civil war. Thus, regimes that repress small protests aggressively are more likely to compromise once larger shocks emerge, while regimes that appear more tolerant of small protests are less willing to share power and repress large protests harshly.

Furthermore, variation in δ due to exogenous historical or geographical factors (e.g. colonial origins) may help to pin down the likelihood of democratic transition or civil conflict: low δ societies generate a "winner takes all" environment that increases the incentives for the regime and the opposition to engage in political violence to maintain power. Since citizens are more likely to falsify their preferences and support whoever that appears stronger, dictators will refuse democratic compromises or power sharing with rival factions if yielding instead of repressing signals weakness. High δ societies are more likely to experience peaceful democratic transitions and avoid civil conflicts.

One unsatisfactory aspect of the analysis is that repression once chosen is fixed and cannot be adjusted even as expectations change. We can consider a dynamic game under which if an equilibrium is not reached after the first repression, then the game is replayed:

1. The regime selects a repression level $r' \in [0, \min\{\hat{\beta}', R\}]$;
2. The citizens update their utilities according to expectations $\{\hat{\alpha}' + r', \hat{\beta}' - r'\}$. We assume that the emotional effect from the last sequence of play remains and the integrity parameter of Type 1 agent is further updated to be $e(r+r')$. If the information mechanism is also permanent, then $r'(1-r)(1-\alpha-\beta)$ of Type 0.5 citizens become Type 1 citizens. The citizens choose y_i according to the utilities.
3. the process is repeated in an analogous manner until the regime regains full support, it democratizes, or Party 1 gains complete power.

If we introduce monetary costs to repression and allow $\hat{\alpha}$ to be the regime's tax base, an alternative to repression may be bribery/social spending or accommodation as considered in Ginkel and Smith (1999) and Passarelli and Tabellini (2017).

Finally, Party 1 does not choose anything in this model: it could impose its own "repression"-revolutionary violence- on supporters of Party 0 in an attempt to maximise $\hat{\beta}$. Bueno de Mesquita (2010) investigated how a "revolutionary vanguard", a relatively extreme opposition group, generates information about anti-regime sentiment by engaging in violence. More successful revolutionary violence informs the public that anti-regime sentiment is higher than they thought, potentially sparking a spontaneous uprising under favorable structural conditions. This extension could greatly enrich the model by showing how under uncertainty as to the popularity of the opposition vis-à-vis the regime, rebel groups may gain or lose momentum by choosing levels of anti-regime violence and specify the conditions under which both parties agree to not fight and allow democratic transition. A wide range of political phenomena could be explained under this framework of combining interrelation between agents' actions and the tradeoffs of political violence.

References

Acemoglu, D., & Robinson, J. A. (2006). *Economic Origins of Dictatorship and Democracy*. New York, NY: Cambridge University Press.

Aytaç, S. E., Schiumerini, L., & Stokes, S. (2018). "Why Do People Join Backlash Protests? Lessons from Turkey". *Journal of Conflict Resolution*, *62*, 1205–1228.

Besley, T., & Persson, T. (2011). "The Logic of Political Violence". *Quarterly Journal of Economics*, *126*(3), 1411–1445.

Bikchandani, S., Hirshleifer, D., & Welsh, I. (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades". *Journal of Political Economy*, *100*(5), 992–1026.

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

Cantoni, D., Yang, D. Y., Yuchtman, N., & Zhang, Y. J. (2019). "Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement". *Quarterly Journal of Economics*, *134*(2), 1021–1077.

Chen, H., & Suen, W. (2016). "Falling Dominoes: A Theory of Rare Events and Crisis Contagion". *American Economic Journal: Microeconomics*, *8*(1), 228-255.

Chwe, M. S.-Y. (2000). "Communication and Coordination in Social Networks". *Review of Economic Studies*, *67*(1), 1–16.

DeNardo, J. (1985). *Power in Numbers: The Political Strategy of Protest and Rebellion*. Princeton University Press.

Edmond, C. (2013). "Information Manipulation, Coordination, and Regime Change". *Review of Economic Studies*, *80*(4), 1422–1458.

Ellis, C. J., & Fender, J. (2010). "Information Cascades and Revolutionary Regime Transitions". *The Economic Journal*, *121*, 763–792.

Ginkel, J., & Smith, A. (1999). "So You Say You Want a Revolution: A Game Theoretic Explanation of Revolution in Repressive Regimes" . *Journal of Conflict Resolution*, *43*(3), 291–316.

Granovetter, M. (1978). "Threshold Models of Collective Behavior". *American Journal of Sociology*, *83*(6), 1420–43.

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

Kuran, T. (1989). "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolutions". *Public Choice*, *61*, 41–74.

Kuran, T. (1991). "The East European Revolution of 1989: Is it Surprising that We Were Surprised?". *American Economic Review, Papers and Proceedings of the Hundred and Third Annual Meeting of the American Economic Association*, *81*(2), 121-125.

Kuran, T. (1995). *Private Truths, Public Lies: the social consequences of preference falsification*. Cambridge, MA: Harvard University Press.

Lohmann, S. (1994). "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-91". *World Politics*, *47*(1), 42–101.

Morris, S., & Shin, H. S. (2002). "Social Value of Public Information". *American Economic Review*, *92*(5), 1521-1534.

Oberschall, A. R. (1994). "Rational Choice in Collective Protests". *Rationality and Society*, *6*(1), 79–100.

O'Halloran, S., Leventoglu, B., & Epstein, D. (2012). "Minorities and Democratization". *Economics and Politics*, *24*(3), 259–278.

Opp, K.-D., & Roehl, W. (1990). "Repression, Micromobilization, and Political Protest". *Social Forces*, *69*(2), 521–547.

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

Pierskalla, J. H. (2010). "Protest, Deterrence, and Escalation: The Strategic Calculus of Government Repression". *Journal of Conflict Resolution*, *54*(1), 117–145.

Rubin, J. (2014). "Centralized Institutions and Cascades". *Journal of Comparative Economics*, *42*(1), 340–357.

Schelling, T. (1978). *Micromotives and Macrobehavior*. New York, NY: W.W. Norton & Company, Inc.

Shadmehr, M. (2015). "Extremism in Revolutionary Movements". *Games and Economic Behavior*, *94*, 97–121.

Shadmehr, M., & Bernhardt, D. (2011). "Collective Action with Uncertain Payoffs: Coordination, Public Signals, and Punishment Dilemmas". *American Political Science Review*, *105*(4), 829–851.

Siegel, D. A. (2011). "When Does Repression Work? Collective Action in Social Networks.". *Journal of Politics*, *73*, 993–1010.

Tullock, G. (1971). "The Paradox of Revolution". *Public Choice*, *11*, 89–99.

Yin, C.-C. (1998). "Equilibria of collective action in different distributions of protest thresholds". *Public Choice*, *97*, 535–567.