The dynamics of revolutions[*]

Moti Michaeli[†]& Daniel Spiro[‡]

March 28, 2017

Abstract

We study the dynamics of revolutions and mass protests. In a unified framework we explain three classes of observed revolutions, two of which are unexplained by earlier models: 1) a revolution initiated by extreme regime opponents dissenting greatly, later joined by moderate dissidents dissenting less; 2) a revolution initiated by moderates dissenting moderately, later joined by extremists; 3) a revolution where extreme regime opponents gradually push the freedom of speech, backed by increased dissent of moderates. These match the dynamics of many major revolutions, e.g., the Iranian Islamic Revolution, the fall of the USSR, the Egyptian Arab Spring and the Tiananmen-Square protests.

Key words: Revolution; Mass protest; Regime; Dissent.

JEL: D74; P26: P5; Z12.

[*] We wish to thank Elias Braunfels, Jean-Paul Carvalho, Sylvain Chassang, Bård Harstad, Arie Melnik, Anirban Mitra, Kalle Moene, Manuel Oechslin, Paolo Piacquadio, Debraj Ray, Kjetil Storesletten, Patrick Testa, Sareh Vosooghi, Yikai Wang, and seminar participants at George Washington University, Oslo Business School, University of Oslo, Tilburg University, King's College, EUI and the ASREC, IMEBESS and NCBEE conferences for valuable comments.

[†] Department of Economics, European University Institute, Italy and the University of Haifa, Israel. Email: motimich@gmail.com.

[‡] Corresponding author, Oslo Business School, Norway. daniel.spiro.ec@gmail.com, Tel: +47 67238461.

1 Introduction

Throughout history, revolutions have led to fast and massive changes in institutional, economic and social environments and, as such, most social-science disciplines have been interested in understanding their causes and dynamics. It is common to divide revolutions against a regime into two categories (Tanter and Midlarsky, 1967). First, *coup detats,* performed by elites or a competing party to the regime.[1] Second, *major revolutions,* driven not by a small group of elites but by popular protest and large social movements. This paper is concerned with the latter category, which includes, e.g., the recent Arab Spring, the toppling of the Shah in Iran in 1978-79, the collapse of the communist regimes in Eastern Europe in 1989 and the protests on Tiananmen Square in Beijing in 1989. In particular, we contribute to the understanding of who will participate in a revolution, which stances these individuals will express and what may spark the revolution.

The workhorse model of revolutions and mass protests is binary.[2] That is, each individual can either support the regime or protest against it, individuals differ in their inclination toward each of these two alternatives and, importantly, the larger the share of individuals that choose an alternative is, the more each individual is inclined to choose so as well. The binary model provides valuable insights on, for instance, thresholds for regime stability. However, since an individual in that model can only choose between complete obedience to the regime and full-blown protest, the binary model invariably predicts that the revolution will be initiated by those who dislike the regime the most. This prediction is inconsistent with the actual evolution of many important revolutions. Consider, for instance, the Arab Spring in Egypt in 2011. These protests were initiated by moderate liberals and moderate conservatives (Lesch, 2011), while those most critical to Mubarak's regime – the Muslim Brotherhood and the Salaffis – were the *last* to join (BBC, 2013; Al Jazeera 2011). Similarly, many of the communist-regime collapses in Eastern Europe in 1989-1991 started by moderates (or regime insiders) (Lohmann, 1994, Przeworski 1991, Breslauer 2002). The binary model is furthermore silent about the *extent* to which each individual will dissent, and how this will change over the course of the revolution. This means that it cannot distinguish between a revolution, such as the Iranian in 1979, where dissent (by Khomeini's followers) was fierce right from the start and later other, more moderate, factions joined (Razi 1987, Moaddel 1992, Ghamari-Tabrizi 2008), and protests such as those on Tiananmen Square in 1989 where the initial dissent was moderate but then got more fierce over time (Zhao 2001). The purpose of this paper is to present a unified framework that accounts for these differences between revolutions and whereby three distinct classes of revolutions and mass

[1] Examples of these are plentiful in both Africa and Latin America and they are typically modeled by assuming the existence of an elite group in society (e.g., Acemoglu and Robinson, 2001).

[2] The binary model was developed by Granovetter (1978) and then discussed and applied in a series of papers by Kuran (1989, 1991, 1995). Other papers using such, or an adjacent, setup are Naylor (1989), Bueno de Mesquita (2010), Olsson-Yaouzis (2012), Edmond (2013), Rubin 2014 and Shadmehr (2015a). See the later literature review for more details.

protests can be explained (a richer account of the examples is provided later in the paper).

1. **A revolution starting with extremists dissenting extremely**. The central characteristic of this class is that those who are most critical to the regime initiate the revolution by dissenting intensely, and gradually less critical individuals join the protests but dissent less than the initiators. This class can be illustrated by the Islamic Revolution against the Shah in Iran in 1978-79 and the fall of the Madero regime in Mexico in 1911-1913.

2. **A revolution starting with moderates dissenting moderately**. The central characteristic of this class is that the revolution starts with moderates expressing moderate critique and, gradually, less moderate individuals join and express harsher critique. This class of revolutions can be illustrated by the Arab Spring in Egypt in 2011 and by many of the communist-regime collapses in Eastern Europe in 1989-1991.

3. **A revolution starting with extremists dissenting moderately**. The central characteristic of this class is that strong opponents of the regime stay in the frontline throughout the protests, while gradually increasing their dissent. This class can be illustrated by the evolution of the protests on the Tiananmen Square in Beijing in 1989 and by the April Revolution that led to the fall of President Rhee in South Korea in 1960.

Our theory provides predictions for which of the three classes of revolutions will occur, and hence which individuals will participate in a revolution at various stages and how extreme their dissent will be. We also provide explanations for other central observations that the binary model cannot account for. For instance, the binary model predicts that only unpopular policies can start a revolution, while popular policies cannot since they ultimately increase the relative attractiveness of supporting the regime. However, history shows that sometimes a popular policy *does* spark a revolution, for instance, Perestroika in the USSR (Brown 1997). That such policies could ignite a revolution came as a surprise to both experts and academics (as documented by, for instance, Kuran 1991 and Lipset and Bence 1994). Following this surprise, Przeworski (1991, p.1) writes that "[a]ny retrospective explanation of the fall of communism must not only account for the historical developments but also identify the theoretical assumptions that prevented us from anticipating these developments". We do just that – by pointing at the limitations of the binary model – and provide predictions for when a popular policy is likely to start a revolution and when not. Another important characteristic of revolutions about which the binary model is silent is whether dissent would come only from one side of the political spectrum or from both sides. Our model accounts for both cases, predicting when dissent will start only on one side (like the radical Muslims in Iran 1979), and when it will start on both sides, with "strange bedfellows" weakening the regime simultaneously (like in Egypt in 2011, where some complained that Mubarak was not sufficiently liberal while others complained he was not sufficiently conservative, or in the USSR in 1987-1991, where hardline Communists criticized the extent of Gorbachev's reforms while others, like Yeltsin, complained the reforms were not sufficiently far reaching).

The theory we develop contains all the core components of the standard binary model with a seemingly simple extension. Rather than having a binary choice between obeying the regime or dissenting against it, an individual can choose the *extent* to which she dissents from a continuum (where not dissenting at all can be interpreted as staying silent or alternatively supporting the regime). The more an individual dissents, the more she will be sanctioned. Just like in the binary model, what makes her possibly dissent despite this sanctioning is that she has a private bliss point (a political view or an economic interest) from which it is costly for her to deviate. Hence, each individual trades off the sanctioning for disobeying the regime against the cost of deviating from her bliss point. These bliss points are heterogenous in the population thus capturing that society consists of different factions. The strength of the regime (i.e., how heavily it sanctions dissent) is endogenous – it decreases in the extent of dissent in society in terms of the number of dissenters and how strongly each one dissents. Then, since the strength of the regime affects the propensity of individuals to dissent, individuals who do not dissent discourage others from dissenting as well – a classic collective-action problem like in most models of revolutions.

The evolution of dissent is shown to depend, to a large extent, on the sanctioning structure the regime is using. A regime that uses a concave sanctioning structure barely differentiates between small and large dissent and hence essentially requires full obedience by the individual to avoid sanctioning. Then, since those who dislike the regime the most perceive the highest cost of obeying it, it will be these individuals – the extremists – who will be first to dissent, and the low marginal punishment once dissenting will push them to express views that deviate greatly from the regime. During the course of the revolution the overall sanctioning power of the regime will gradually fall, and less extreme individuals with less extreme dissent will join too. This fits the pattern of the first class of revolutions (illustrated by the Islamic revolution in Iran in 1979). In contrast, the use of a convex sanctioning structure by the regime means that small dissent is not so costly while large dissent is very costly. Hence, under such sanctioning, no one dissents a lot. During a revolution, the sanctioning gets gradually weaker and hence the maximal dissent increases over the course of the revolution – the freedom of speech is pushed further. This fits the pattern of the second and third classes of revolutions (illustrated by the Arab Spring in Egypt 2011 and the protests on Tiananmen square in 1989 respectively).[3]

Who in society will be dissenting first (thereby starting the revolution) largely depends on the cost of deviating from one's ideological (or economic) bliss point. In a society that

[3] Strictly speaking, the sanctioning does not necessarily need to be convex for dissent to start moderately. It is sufficient that the regime's sanctioning is not too concave.

is characterized by individuals with a convex cost, individuals will find it easy to deviate slightly from their bliss points, while large deviations will be very costly. Hence, in such societies, individuals with private views close to the regime will tend to obey it, while individuals whose private views are very far from the regime's policy will dissent more. Consequently, the most extreme types will be the ones dissenting the most and thus leading the way during the revolution. This explains the pattern of the first (Iran) and third (Tiananmen square) classes of revolutions. In contrast, in a society that is characterized by individuals with a concave cost of deviating from their bliss points, individuals will find it very costly to deviate even a little from their bliss points, but deviating more will be only marginally more costly. Hence, if they do deviate, they might as well align with the regime, for instance by remaining silent. Those who will find it the hardest to express their private views and hence are prone to stay silent are the extremists, because their views are sanctioned more than the views of moderates. Thus, they will be the ones aligning with the regime. This means that the most deviant expressions will be stated by moderates, who are thus the ones initiating the revolution. During the revolutionary process, as the regime's sanctions get weaker, extremer types will find it possible to express their private views and thus start dissenting. This fits the pattern of the second class of revolutions (Egypt), where the most extreme regime opponents are the last ones to join the revolution and the revolution is initiated by moderate forces (or party insiders like in the USSR).

Who starts the revolution and how dissent evolves during the revolution has implications for whether popular or unpopular policies will trigger a revolution; when a revolution is more likely to be one-sided and when two-sided; at what stage a revolution is most likely to fail and what the regime can do to achieve it; and when the revolutionary momentum will mainly be driven by new recruits versus by gradual increases of dissent of current participants. We analyze these questions in the paper. Following a brief literature review, section 2 outlines the model and presents the main results. Sections 3-5 analyze in depth the three classes of revolutions, each in turn, and provide more details on the historic examples briefly discussed earlier. Section 6 provides empirical predictions and Section 7 concludes.

1.1 Relation to previous research

Unlike this paper, most of the previous theoretical literature on major revolutions and mass protests utilizes a binary model (see, for instance, Granovetter 1978; Kuran 1989; Naylor 1989; Angeletos et al. 2007; Olsson-Yaouzis 2012; Edmond 2013; and Rubin 2014).[4] As explained, such models cannot account for revolutions that are initiated by moderates or, more generally, with moderate dissent, nor for how dissent evolves during the revolution or for revolutions triggered by popular policies.

More broadly, however, obeying a regime and conforming to a social norm are theo-

[4] Note that in Rubin's (2014) paper, although the regime can choose its policy from a continuum, the individual has a binary decision whether to support or not support the regime.

retically quite similar and in the literature on social norms some non-binary models exist (Bernheim 1994, Kuran and Sandholm 2008, Manski and Mayshar 2003, Michaeli and Spiro 2015,2017). The structure of the individual trade off – between obeying a regime or norm and following one's heart – is thus shared by these papers and the current paper. However, Bernheim (1994), Manski and Mayshar (2003) and Michaeli and Spiro (2015) are concerned with static equilibria hence are silent about the dynamics which obviously are central in revolutions, Kuran and Sandholm (2008) analyze integration between groups and no regime exists in their framework, and Michaeli and Spiro (2017) study the existence of a social norm in equilibrium.

Granovetter (1978), Kuran (1989) and many subsequent papers offer a dynamic setting in which individuals take the actions of others as given. This kind of analysis, which abstracts from strategic considerations on the individual level, seems adequate for analyzing major revolutions and mass protests whereby, literally, millions of individuals may participate. We therefore adopt this approach in our paper too.[5] Strategic (or more precisely, informational) considerations by revolutionary participants are analyzed by Angeletos et al. (2007) and Edmond (2013), but in a binary framework hence with limited predictability for the issues we are interested in. Strategic behavior of revolutionary leaders has been analyzed by Bueno de Mesquita (2010) and Shadmehr (2015a), where the latter is the paper most related to ours. In their papers, citizens have a binary choice between supporting the regime or following a protest leader who offers an alternative policy to the regime while taking into account the support she will get by the population. Hence the focus of these papers is different than ours. The similarity of Shadmehr's (2015a) paper to ours is in that the regime's punishment structure affects choices. However, his model is limited like the binary model in predicting that those with extreme views are in all scenarios part of the revolution, while those sufficiently close to the regime are never part of the revolution. As we exemplify with Egypt and Eastern Europe, this prediction is not true in many important cases. Furthermore, with our focus on the dynamics of the revolutions, we answer questions (not analyzed by Shadmehr 2015a) such as how the statements will evolve over time, who will join the revolution at what point in time and which regime policies (popular or unpopular) will trigger a revolution.

Our paper is also related to the literature on the collective-action problem (see Olson 1971 and Tullock 1971 for early treatments and, e.g., Oliver & Marwell, 1988; Esteban, 2001; Esteban & Ray, 2001 for more recent work) by the fact that, in our model, individuals who do not dissent discourage others from dissenting as well, implying that agents dissent too little. What makes an individual *partly* overcome the problem and dissent against the regime is that she wants to express her own ideology – an aspect that has been identified as an important determinant of revolutions (Goldstone, 2001). We do not, however, directly

[5] We have also solved a version of our model with a small number of strategic agents and the main results remain the same. Such a model is, however, not tractable for obtaining our further results.

contribute to the full solution to collective-action problem nor to which type of groups are more likely to overcome it (like, for instance, Esteban 2001 does) but we do provide novel predictions on what kind of exogenous changes in policy enable the population to overcome the problem.

2 The model

We start by describing a static version of the model and then add a dynamic structure to it. Society consists of a continuum of infinitesimal individuals of unit mass and of a political regime. The regime has a policy $R \in [-1, 1]$ which can be thought of as a point on a left-to-right political scale. Focusing on revolutions and mass protests against a given regime, we let R be exogenous (capturing the regime's ideology). Each individual takes a stance $s \in \mathbb{R}$. The regime sanctions stances that deviate from its policy ($s \neq R$), with larger deviations representing harsher critique of the regime, which in turn is sanctioned more heavily. $s = R$ can be interpreted as the individual staying silent. The sanctioning is represented by the following *punishment* function:

$$P(s, R, K) = K |s - R|^{\beta} \ , \ \beta > 0. \tag{1}$$

The overall severity of punishment (sanctioning), represented by the parameter K in (1), is endogenous and depends on the aggregate dissent in society. Let S denote a distribution of stances taken by the individuals in society. Now suppose one individual changes her stance from s to some s' and denote this new distribution of stances by S'. The *approval* of the regime, denoted by A, has the following properties.

$$A \in [0, 1] : \ A(S) \geq A(S') \ \text{ iff } \ |s - R| \leq |s' - R| \, . \tag{2}$$

That is, the approval of the regime is decreasing the more dissenting each individual stance is. The overall severity of punishment K, to which we also refer as the *strength* of the regime, is proportional to A,

$$K = \bar{K} A$$

where $\bar{K}$ is a parameter capturing the *force* of the regime. K is endogenously capturing the actual strength of the regime, so that the more approving the population is of it, the easier it is for the regime to punish dissenters. $\bar{K}$ could represent, for instance, the per capita law-enforcement forces used by the regime to sanction dissent. Then, the increase of K in A could mean that the more dissent there is, the less likely it is for an individual dissenter to get caught. Or alternatively, A could stand for the proportion of the law enforcement forces that stay loyal to the regime when asked to use force against dissenting civilians. β captures the curvature of the sanctioning system, which will be important for the analysis. A regime with a large β (> 1) uses convex sanctioning and hence is tolerant towards critique as long as it is not too deviant. A regime with a small β (< 1) uses concave sanctioning whereby it punishes rather heavily even small dissent but does not distinguish much between small and large dissent.

Each individual has a privately preferred political policy or opinion $t \in T \subset \mathbb{R}$, also referred to as the individual's bliss point or type. Let $g(t)$ denote the probability-density function of types and assume it is continuous. When expressing a stance s, the individual bears a cost for deviating from her bliss point:

$$D(s,t) = |s - t|^{\alpha} \ , \ \alpha > 0. \tag{3}$$

D can be interpreted as discomfort from expressing a political opinion not in line with a person's conviction (or, if t reflects one's optimal economic behavior, as a material cost of deviating from the person's economic interests). The choice problem of an individual with $t \neq R$ is how to trade off the sanctioning when dissenting against the regime and the disutility of deviating from her own privately held opinion. That is, the individual minimizes

$$L(s; t, R, K(S)) = D(s,t) + P(s, R, K(S)). \tag{4}$$

It is immediate from this choice problem that the individual will take a stance somewhere weakly in between t and R. The extent to which the individual feels forced to go towards the regime depends on the regime's strength $K(S)$ and hence indirectly on the stances taken by all individuals in society.

Being interested in how a regime's strength interacts over time with the behavior of individuals in society, we will now add a simple dynamic structure to the model. These dynamics are standard in games with large populations (e.g., Young 1993; Kaniovski et al. 2000; Young 2015) and in the analysis of revolutions (e.g., Granovetter 1978; Kuran 1989). The stances of individuals in period $i+1$ are determined by a mapping from the type space to the stance space $s^*_{i+1} : T \rightarrow \mathbb{R}$, such that,

$$s^*_{i+1}(t, R, K(s^*_i)) = \arg\min_{s_{i+1}} \{L(s_{i+1}; t, R, K(s^*_i))\}. \tag{5}$$

This formulation means that the regime strength (K) that affects stances in period $i+1$ is determined by the stances taken in period i. This implies that s^*_{i+1} is a function of s^*_i and that

$$A_{i+1} = f(A_i)$$

where f describes the dynamics of approval between periods. This dynamic structure could, for example, represent the loyalty or assertion of the regime's troops at day $i+1$ of a revolution after observing the dissent of the population on the previous day. We wish to emphasize that we choose these adaptive dynamics for tractability and for brevity in

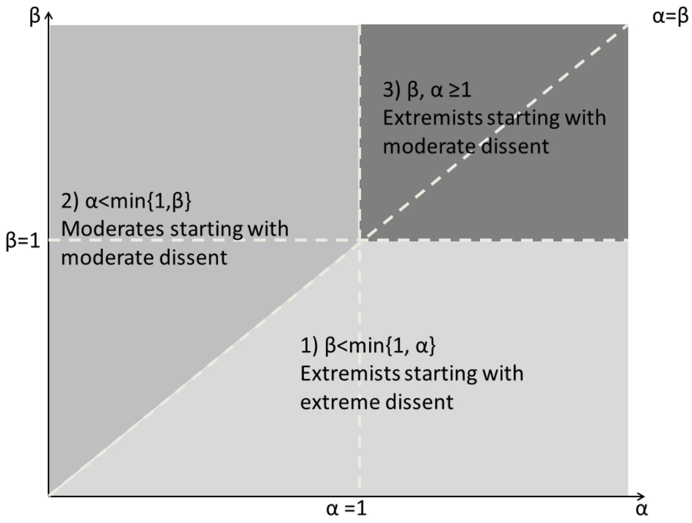


Figure 1: Parameter space of the different classes of revolutions.

presenting the results but the specific dynamic modeling does not drive our results.[6] A steady state (which is equivalent to a Nash equilibrium in the static model) is achieved when $s_{i+1}^{*}\left(t, R, K\left(s_{i}^{*}\right)\right)=s_{i}^{*}\ \forall t$, which also yields $f\left(A_{i}\right)=A_{i}$. We set $A_{i}=0$ when $s_{i}^{*}(t)=t\ \ \forall t$. This is always a steady state as it implies that $P(s, R, K)=P(s, R, 0)=0$ and so $s_{i+1}^{*}(t)=t\ \ \forall t$. It can be interpreted as a state with no regime or as a state with a regime that has no control over the population. We will consider a steady state to be stable, with its approval denoted by A_{ss}, if there is convergence back to it following a small perturbation to A_{ss}. Otherwise the steady state is unstable, with its approval denoted by A_{uss}. Our measure for the stability of the regime following a shock to its approval is the distance between the regime's approval at a steady state A_{ss} and the approval at the closest unstable steady state below it, i.e., $A_{ss}-A_{uss}$, because the zone of convergence to A_{ss} from below is $A \in\left(A_{uss}, A_{ss}\right)$. A revolution is defined as a dynamic process where the approval is converging to a new, lower, steady state (i.e., a revolution is not a situation where a small change to a parameter leads to a small change in the steady-state approval). A successful revolution is one where $A=0$ in the new steady state and a failed revolution is one where $A>0$ in the new steady state. Catalytic events are events that may trigger a revolution. These are exogenous changes or shocks that either imply that a previously stable steady state seizes to exist or decrease the approval to a point where the approval will, endogenously, decrease further.

The main focus of our analysis is on the evolution of participation (i.e., which types dissent) and of statements (i.e., which stances they express) over time during the revolution.

[6] We have also solved a version of the model with forward-looking agents and strategic interaction between the agents. The main results about the three classes of revolutions are the same but it is substantially more complicated to show our further results about two-sidedness, policies that initiate a revolution and failed revolutions.

For short, we will refer to individuals with private views far from the regime (large $|t-R|$) as extremists and to those with private views close to the regime (small $|t-R|$) as moderates. That is, a type's extremeness is always relative to the regime – a liberal democrat under the Taliban regime is an extremist in our definition. The model predicts three classes of revolutions depending on the combination of the parameters β and α, as depicted in Figure 1 and expressed in the following proposition.[7]

Proposition 1 *There are three exhaustive classes of revolutions:*

1. *(**Extremists starting with extreme dissent**) If $\beta < \min\{1, \alpha\}$, then initially the dissenters are extremists, and later in the revolution types who are more moderate join, but dissent less than the initial extremists.*

2. *(**Moderates starting with moderate dissent**) If $\alpha < \min\{1, \beta\}$, then initially the dissenters are moderates, and later in the revolution types who are more extreme join and dissent more than the initial moderates.*

3. *(**Extremists starting with moderate dissent**) If $\beta \geq 1$ and $\alpha \geq 1$ (and at least one inequality is strict), then throughout the revolution the most dissenting types are extremists, who start by dissenting moderately and increase their dissent over time.*

These three classes of revolutions fundamentally differ in how participation and statements evolve during the revolution. The first is a revolution that starts with extremists voicing extreme critique and gradually recruiting individuals who are more moderate. The second is a revolution in which moderates start voicing their mild critique of the regime, gradually making extremer individuals stand up for their (extremer) views as well. The third is a revolution in which the extremists constantly express the most deviant views but start with moderate dissent and gradually push the freedom of speech by dissenting more and more over time.

We defer explaining the intuition of Proposition 1 to the upcoming three sections, where we analyze each class of revolutions separately and obtain further results. Most of the results in the paper, and in particular those stated in Proposition 1, hold for the very general formulation of the approval function A in (2) and any continuous distribution of types $g(t)$, and in fact can be derived also with more general functional forms for P and D. However, showing some specific further results requires a more explicit functional form of A and of the distribution of types. For analytical tractability we will assume from now onwards that $t \sim U(-1,1)$ and that the approval of the regime is linear in the deviations from it. That is,

$$A = \max\left\{0, 1-\lambda \int_{-1}^{1} |s(t)-R|\, dt\right\}. \tag{6}$$

[7] For brevity we ignore here the special case of $\alpha = \beta \leq 1$ with its unique technicalities.

This is a special case of (2) where $A = 1$ if nobody dissents ($s(t) = R\ \forall t$). We normalize $\lambda = 1$ so that a non-biased regime ($R = E[t] = 0$) has zero approval precisely when all types speak their minds ($s(t) = t\ \forall t$).[8] This ensures that for *any* $R \in [-1, 1]$ we get $A = 0$ whenever all types speak their minds. It further implies that, when a regime is biased ($R \neq 0$), A may equal 0 also without all types speaking their minds (which reflects that, when all speak their minds under a biased regime, dissent is larger than when all speak their minds under a non-biased regime). This normalization is largely without consequence apart from implying that even a central regime loses all of its strength when all speak their minds (which would not be true for $\lambda < 1$). Throughout the paper we refer to a regime with $R \neq 0$ as biased and use $|R|$ as a measure of the regime's bias.

3 Extremists starting with extreme dissent

3.1 Analysis

We start by considering the case where $\beta < \min\{\alpha, 1\}$. This case can be further divided into two subcases: $\beta < \alpha \leq 1$ and $\beta < 1 < \alpha$. While these two cases differ in some details, they are largely the same from the point of view of what we are interested in. Hence, for brevity, we will focus here on the subcase $\beta < \alpha \leq 1$.[9]

We begin by analyzing dissent within a time period. By differentiating L twice with respect to s it is immediate that when $\beta < \alpha \leq 1$ the second-order condition is not fulfilled, implying that an individual will choose either $s(t) = R$ or $s(t) = t$. It is simple to further show that there exists a cutoff distance $\Delta = K^{\frac{1}{\alpha-\beta}}$ such that all types closer to the regime than Δ will fully follow the regime (i.e., $s(t) = R$ when $|t - R| \leq \Delta$) while types further from the regime than Δ will speak their minds (i.e., $s(t) = t$ when $|t - R| > \Delta$), as illustrated for a biased regime in Figure 2. Hence, the regime induces silence by those who largely agree with it. The intuition is easy to understand. The important property of this case is that β is relatively small, which implies that the regime applies a (very) concave punishment whereby even small dissent is heavily punished while more extreme dissent is punished only slightly more. This will induce an individual to either fully follow the regime or, if she does not fully follow it, she may as well dissent quite a lot. Then, since types far from the regime perceive the highest cost of discomfort from following the regime, these types will be the ones who may dissent – and dissent quite a lot if they do given their extreme views – while types close to the regime will be silent. The cutoff Δ between those following the regime and those who do not (if they exist) is naturally increasing in the strength of the regime K, so that a stronger regime sees less dissent.

The result that extremists speak their minds and moderates are completely supporting the regime has important implications for the stability of regimes and for the revolutionary

[8] I.e., $\lambda = 1/\left(2\int_0^1 t dt\right) = 1$.

[9] See sections A.2.2 and A.3.1 in the appendix for a treatment of the other subcase ($\beta < 1 < \alpha$).

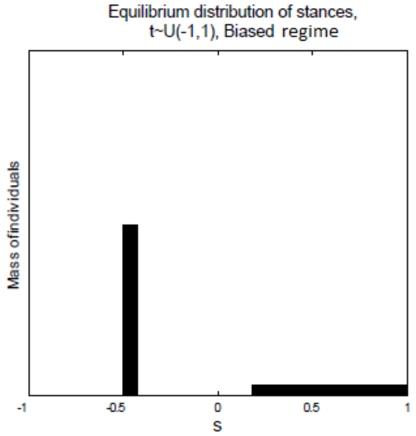


Figure 2: An illustration of an equilibrium distribution of stances under a biased regime for $\beta < \alpha \leq 1$.

dynamics as expressed in the following proposition.

Proposition 2 *When $\beta < \alpha \leq 1$:*

1. ***Existence of a stable steady state:*** *A stable regime exists iff it employs sufficient force, and the more biased its policy is the more force it needs to employ.*
2. ***Catalytic events:*** *A revolution may start following a shock to the regime's approval or force or following implementation of unpopular policies.*
3. ***Dynamics of participation:***
   (a) *Initially only the most extreme types participate in the revolution, but over time types who are more moderate join it too.*
   (b) *For any regime with $|R| \neq 0$, the revolution will start only on one side of the political spectrum.*[10]
4. ***Dynamics of statements:*** *Initially dissents are extreme and over time, as moderates join the revolution, the new statements are more moderate.*

We start by explaining the dynamics of the revolution (parts 3 and 4) since this largely determines what makes a regime stable and which events may initiate a revolution. The revolutionary process follows from the dynamics of the cutoff between those who dissent and those who do not (Δ_i). As explained earlier, when the regime uses (very) concave sanctioning, it induces dissent by extremists but not by moderates. This means that, if a revolution starts, the first ones to dissent are the most extreme types. When these

[10] Unless there is a very large shock to the force or approval of the regime or a very large change to its policy.

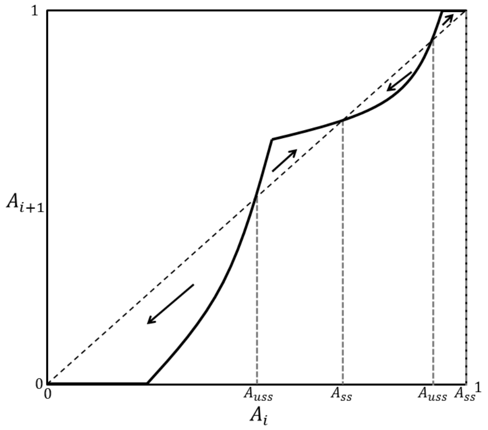


Figure 3: Stylized phase diagram for the case $\beta < \alpha \leq 1$ and a moderately biased regime ($|R|\,]0, 0.5[$). The full line depicts the intertemporal-dynamics function $A_{i+1} = f(A_i)$ and the dashed line depicts the 45-degree line where $A_{i+1} = A_i$. The vertical lines depict the stable (A_{ss}) and the unstable (A_{uss}) steady states.

extremists start dissenting, the strength of the regime falls, which makes it possible also for less extreme types to dissent. This way, increasingly moderate types join the revolution and they dissent less than those who started it (as summarized in parts 3a and 4 of the proposition). If the regime is, say, left of center ($R < 0$), the first dissenters will be on the far right − the revolution starts only on one side (part 3b of the proposition, illustrated in Figure 2). During this phase, the revolutionary momentum is rather low, since new recruits are only coming from the right, while later, if the regime has gotten sufficiently weak, new recruits might appear also on the left side of the regime. This has important implications for the fragility and success of a revolution as will be explained later.

As a tool to understand the additional results, consider the phase diagram in Figure 3 which depicts a stylized example of the intertemporal-dynamics function $A_{i+1} = f(A_i)$ for a moderately biased regime ($|R| \in\,]0, 0.5[$). The higher is current approval (A_i), the higher is the regime's strength (K_i), which implies less dissent thus higher approval in the next period (A_{i+1}). Hence, A_{i+1} is a (weakly) increasing function of A_i as can be seen in the figure. Quite naturally, for any approval level A_i, an increase of the regime's force $\bar{K}$ raises A_{i+1} (through an increase of $K_{i+1} = \bar{K}A_i$). This implies that the function $A_{i+1} = f(A_i)$ in Figure 3 shifts up as $\bar{K}$ is increased. For sufficiently small $\bar{K}$, A_{i+1} is always below the 45-degree line, implying no stable regime exists. This naturally implies that there is a minimum amount of force a regime has to employ to stay stable. For sufficiently large $\bar{K}$ there may exist one, two, or three (inner) intersections with the 45-degree line. The first intersection from the left is an unstable steady state, the second is stable and the third is unstable. Additionally, as is the case in the figure, there is one stable steady state at $A_i = 0$ and there may be one at $A_i = 1$.

As those with views far from the regime are the ones dissenting, a biased regime, with policies far from most of the population's views, will induce more dissent for any given level of regime strength K_i (biasness shifts down the dynamic approval function in Figure 3). Hence, biasness has to be compensated for by the employment of more force ($\bar{K}$) for a stable steady state to exist (part 1 of the proposition).[11] Furthermore, since increasing the bias decreases the approval function, the unstable steady states move right while the stable steady states move left in the phase diagram. This means that biased regimes are inherently less stable and that an implementation of unpopular policies (increase in the bias) may ultimately be the catalytic event that starts a revolution by making a previously stable steady state seize to exist (as stated in part 2). Other catalytic events are a shock to the regime's force (a shift downward of the approval function) or a shock to the approval (lowering A_i), if they put $A_{i+1} = f(A_i)$ in a zone where the political system converges downwards.

The further properties of the phase diagram depicted in Figure 3 (for $|R| \in]0, 0.5[$) is that the function $A_{i+1} = f(A_i)$ is first flat near zero (unless $R = 0$), then rises convexly, kinks downwards and then rises convexly again. The flat initial part is where current approval is so low that the regime will not be able to gain any approval at the next period ($f(A_i) = 0$). To see why a kink exists, consider a left-biased regime. When approval A_i is low, there will be dissent on both sides of the regime in the next period since Δ_{i+1} (the range of moderates not dissenting) will be small. As A_i increases, the dissent falls on *both* sides of the regime implying A_{i+1} is a steep function of A_i. The kink is the point where A_i induces $\Delta_{i+1} = 1 - |R|$ (in Figure 2, $1 - |R|$ is the distance from the regime to the left edge corner). After this point, there is no way to further decrease dissent on the left side of the regime. From here onwards an increase in A_i will reduce dissent only on the right side of the regime, as illustrated in Figure 2, which implies that A_{i+1} becomes less steep.

As explained earlier, when the kink exists in the phase diagram there may be up to two stable steady states with $A_{ss} > 0$ (one internal and one where $A_{ss} = 1$ as in Figure 3). If there is only one stable steady state with $A_{ss} > 0$, a shock that eliminates it will lead to a succesful revolution (one that ends with $A_{ss} = 0$). However, when two stable steady states with $A_{ss} > 0$ exist, a revolution that starts from $A_{ss} = 1$ may fail to topple the regime. Suppose the regime is left biased. The revolution will start with the extremists on the right side recruiting less extreme followers on their side of the political scale. But, since new recruits come only from the right side of the regime, the momentum of the revolution will be low and the revolution fragile. In particular, the revolution is bound to fail if the shock that sparks it eliminates only the stable steady state with $A_{ss} = 1$, while the internal steady

[11] More precisely, an increase in bias shifts the convex part to the right of the kink downwards and at the same time widens this part outwards in both directions. This has to do with the fact that biasness affects dissent not through affecting Δ – which is independent of $|R|$ – but through affecting the actual mass of types at distance larger than Δ from the regime, which increases in $|R|$ when the regime is sufficiently biased to induce dissent only at the opposite extreme side, as can be seen in Figure 2.

state is not eliminated. However, if the revolution eventually reaches the stage (to the left of the kink) where approval is so low that also the most extreme leftists start dissenting, then the revolution becomes two-sided, gains momentum and is bound to succeed. Thus, a revolution that starts with extreme dissent is fragile initially but strong at later stages. When the regime is instead very biased ($|R| \in [0.5, 1]$), the phase diagram will not contain the left convex part. This implies that the regime can collapse even if dissent is only on one side of it, reflecting the weakness of very biased regimes. For example, a very left-biased regime can be toppled by a purely right-wing revolution.

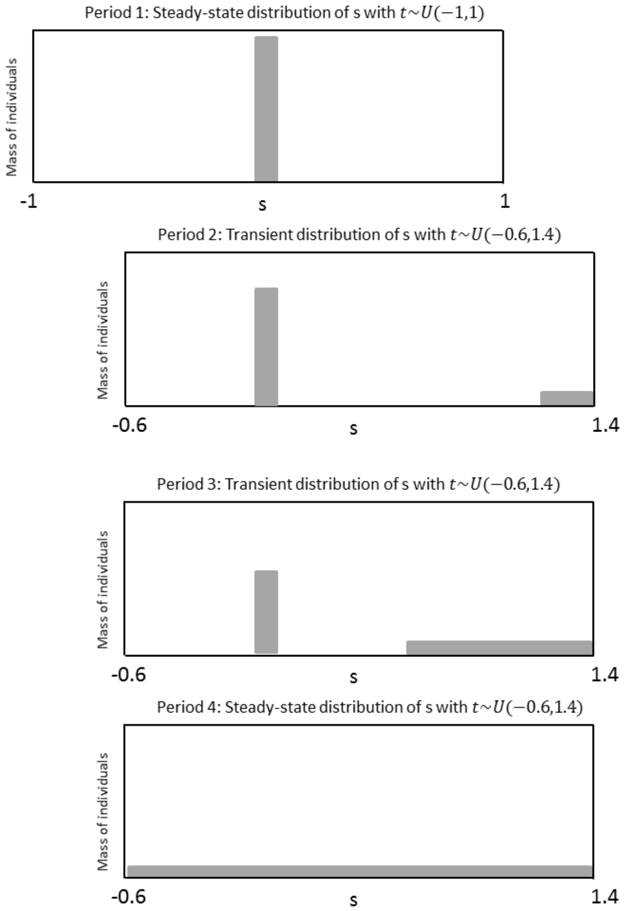
In order to prevent the success of the revolution, the regime has to either increase its force ($\bar{K}$) or implement policies that are more popular (thereby lifting the dynamic approval function). It is further worth noting that a shift of private opinions, say to the right, is equivalent to the regime changing its policies to the left. This is since it is the relative position of the type space vis-à-vis R that matters. This means that our results about a change in the regime's policy have an equivalent in an opposite change of private preferences. For instance, what may start a revolution is that, over time, the private preferences of the population shift away from the regime's policies as depicted in Figure 4.

3.2 Historic examples

We will now provide examples of actual revolutions that fit the dynamics described in this section – of extremists starting with extreme dissent. Note that this is not an empirical test of the model. Rather the purpose is to illustrate that the patterns obtained theoretically have been observed in reality. A proper empirical test would link the parameter conditions (in particular for β and α) of a large set of revolutions to the predictions of the model. This is in principle possible but requires gathering a large set of detailed data which is beyond the scope of this paper. In Section 6 we provide a set of testable predictions (based on the results found here and in the upcoming sections) that could be used for a proper empirical test of the model and briefly discuss what kind of data is needed and how to get around problems of unobservable parameters.

The overall pattern of the class of revolutions just described aligns with the dynamics of the Iranian Islamic Revolution in 1978-79. This revolution began following a gradual increase in the misalignment between the Shah's secular policies and the increasingly religious sentiments in society (Moaddel, 1992). In line with Figure 4, following this misalignment the revolution was initiated by the hardest opponents of the regime, i.e., by Khomeini and his closest group, who held an extreme religious ideology and started dissenting extremely (violently).[12] Then gradually more moderate individuals joined the revolution (Razi 1987, Moaddel 1992, Ghamari-Tabrizi 2008, Shadmehr 2015b). These moderate individuals, while

[12] It is very hard to pin down the exact starting point of the revolution. Some trace it back to 1977 and even earlier. Our own account starts with the first violent clashes at Qom in January 1978 and the subsequent riots in several cities in February that year. The initiators of these events where the followers of Khomeini.



Period 1: Steady-state distribution of s with $t \sim U(-1,1)$

Period 2: Transient distribution of s with $t \sim U(-0.6,1.4)$

Period 3: Transient distribution of s with $t \sim U(-0.6,1.4)$

Period 4: Steady-state distribution of s with $t \sim U(-0.6,1.4)$

Figure 4: Distribution of stances over time in a stylized case of a revolution starting with extremists dissenting extremely ($\beta < \alpha \leq 1$). $R = 0$ and fixed while the distribution of types changes.

being part of the revolution, advocated less extreme policies and used less extreme slogans than Khomeini and even among Khomeini's closest supporters many were advocating a less religious policy than Khomeini (Ghamari-Tabrizi, 2008). During this process, the weakening of the punishment on dissent came in the form of removal of censorship and increased usage of televised debates, i.e., an acceptance of dissent in general (Milani 1994, p.117). In line with our theoretical results it has been claimed that in order to remain in power, the Shah "either had to crush the growing movement or to relinquish some of his power and strike a deal with the moderate faction of the popular movement. He opted to do neither" (Milani 1994, p.116). That is, what could have possibly saved the Shah was either an increase of force ($\bar{K}$ in our model) or the implementation of popular policies (moving R in the religious direction).

Another example for this class of revolutions is the Mexican Revolution of 1911-1913 against Francisco Madero.[13] Madero was formally and democratically elected president in October 1911. In the Mexican context Madero can be considered a centrist. Shortly after being elected, opposition to Madero's regime increased from both the right-wing conservatives, who saw him as too liberal, and from leftist "extremists" (former revolutionary fighters), who saw him as too conservative.[14] As unrest propagated in the country, more moderate factions joined the protests. Some of them leaned to the left, such as those led by revolutionary general Pascual Orozco, who advocated social reforms that called for better working hours, pay, and conditions. Other leaned to the right, such as the supporters of General Bernardo Reyes, who used to be a minister in the previous conservative government (Garner 2001, p. 209). The Madero presidency was unraveling and its approval continued to deteriorate, eventually even among his political supporters in congress, the so-called Renovadores (Katz, 1998).

The revolution against Madero was two-sided already early on as both leftists and rightists started protesting essentially simultaneously. This is consistent with our model as Madero was a centrist. Our model further predicts that once the revolution is two-sided it cannot fail unless either 1) the regime takes harsh measures against the protests, which Madero did not do; or 2) the regime implements popular policies, which Madero could not do already being a centrist and hence any policy would have alienated either the leftists or the rightists even more.[15] In February 1913, only fifteen months after being inaugurated as

[13] The broader Mexican Revolution was in fact a chain of revolutionary events occurring in the 1910s, and this specific stage of the revolution is the second in a series of regime changes. We are referring here to the protests and revolution against the regime of Francisco Madero, which culminated in the "Ten Tragic Days" (9–19 of February 1913).

[14] Madero was considered too leftist by the conservatives since he came to power with the support of the left, most notably Emiliano Zapata. At the same time Madero was considered too conservative by the leftists since he did not want to implement a massive land reform of breaking up large estates.

[15] Madero's allies in congress, the Renovadores (the "renewers") criticized Madero for making "compromises and concessions to the supporters of the old regime" (Katz 1998, pp. 196-97). Even if the Renovadores were right in describing his policies as trying to peas the conservatives, this was clearly not enough to save the regime and was maybe even counterproductive given the widespread leftist sentiments in the Mexican society.

president, Madero was forced to resign and shortly afterwards he was murdered.

4 Moderates starting with moderate dissent

4.1 Analysis

We move now to the case where $\alpha < \min\{\beta, 1\}$. This case can be further divided into two subcases: $\alpha < \beta \leq 1$ and $\alpha < 1 < \beta$. While these two cases differ in some details, they are largely the same from the point of view of what we are interested in. Hence, for brevity, we will focus here on the subcase $\alpha < \beta \leq 1$.[16]

As in the previous section we start by describing dissent within a time period. By differentiating L twice with respect to s it is immediate that when $\alpha < \beta \leq 1$ the second-order condition is not fulfilled, implying that an individual will choose one of the corner solutions: either $s(t) = R$ or $s(t) = t$. It is simple to further show that there exists a cutoff distance $\Delta = K^{\frac{1}{\alpha - \beta}}$ such that all types further from the regime than Δ will fully follow the regime while types closer to the regime than Δ will speak their minds, as illustrated for a biased regime in Figure 5. Hence, as opposed to the first class of revolutions, here the regime induces silence by those who dislike it the most while those who largely agree with the regime pose mild critique of it. To understand the intuition behind this result note that the important property of this case is that α is relatively small and in particular smaller than β. Consider, for instance, the special case of $\beta = 1$. First note that, since $\alpha < 1$, individuals will perceive a relatively high cost from even a small deviation from their bliss points. Hence, they will either speak their minds or, if this is too difficult given the punishment, be willing to go a long way to avoid punishment by the regime. Then, as $\beta = 1$ implies that speaking one's mind is considerably harder for extremists, they will be the ones submitting to the pressure and following the regime, while moderates will find it bearable to speak their minds. The cutoff value Δ – between those speaking their minds and those staying silent – is decreasing in K, reflecting that the stronger the regime is, the smaller is the share of the population speaking their minds. The result that extremists keep silent while the moderates are speaking their minds has important implications for the stability of regimes and for the revolutionary dynamics as expressed in the following proposition.

Proposition 3 *When $\alpha < \beta \leq 1$:*

1. ***Existence of a stable steady state:*** *A stable regime exists iff it employs sufficient force, and the more biased its policy is the less force it needs to employ.*

2. ***Catalytic events:*** *A revolution may start following a shock to the regime's approval or force or following implementation of popular policies.*

3. ***Dynamics of participation:***

[16] See sections A.2.3 and A.3.2 in the appendix for a treatment of the other subcase ($\beta < 1 < \alpha$).

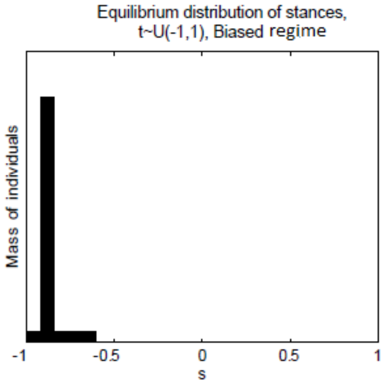


Figure 5: An illustration of an equilibrium distribution of stances under a biased regime for $\alpha < \beta \leq 1$.

(a) *Initially only the most moderate types participate in the revolution, but over time types who are more extreme join it too.*

(b) *For any regime with $|R| \neq 1$ the revolution will be two-sided throughout.*

4. ***Dynamics of statements****: Initially dissents are moderate and over time, as extremists join the revolution, the new statements are more extreme.*

We start by explaining the dynamics of the revolutions (parts 3 and 4 of the proposition) since this largely determines what makes a regime stable and which events may initiate a revolution. The revolutionary process follows from the dynamics of the cutoff between those who dissent and those who do not (Δ_i). As explained earlier, when individuals perceive a (very) concave cost of deviating from their blisspoints, it induces dissent by moderates but not by extremists. This means that if a revolution starts, the first ones to dissent are types with bliss points close to the regime and they will dissent only moderately (as illustrated in Figure 5). Strictly speaking, those closest to the regime speak out already before the revolution and, when the revolution starts, dissent starts also by types slightly more critical of the regime. When these types dissent, the strength of the regime falls, which makes it possible also for extremer types to dissent and express their extremer views publicly. This way, both the participants and their expressed views are growing more extreme over time (as summarized in parts 3a and 4 of the proposition). This further implies that for any regime (except for the most biased regimes, in which $|R| = 1$) the dissent will be two-sided right from the start (part 3b of the proposition) – some will be complaining that the regime is too leftist and some that it is too rightist.

To understand the additional results, consider the intertemporal-dynamics function $A_{i+1} = f(A_i)$ depicted in Figure 6. It is first flat and then starts to increase concavely. $A_{i+1} = f(A_i)$ then kinks upwards at some point (provided that $|R| \neq 0$) and is concave thereafter. The presence of the kink has important implications for the potential revolution.

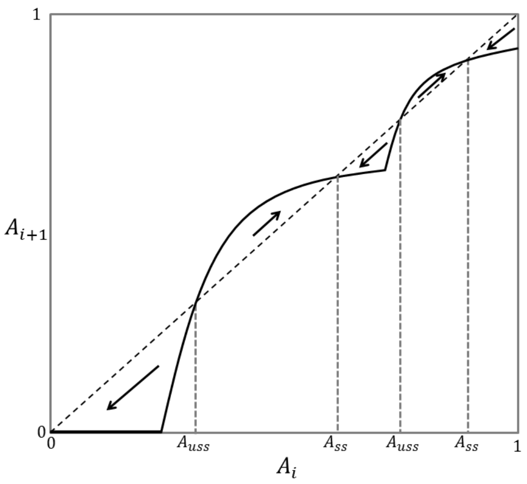


Figure 6: Stylized phase diagram for the case $\alpha < \beta \leq 1$ and a biased regime. The full line depicts the intertemporal-dynamics function $A_{i+1} = f(A_i)$ and the dashed line depicts the 45-degree line where $A_{i+1} = A_i$. The vertical lines depict the stable (A_{ss}) and unstable (A_{uss}) steady states.

To see why this kink exists, recall that the cutoff between those who speak their minds and those who obey the regime (Δ) is large for a small K. Consider now a left-biased regime (as the one depicted in Figure 5). The flat initial part of $A_{i+1} = f(A_i)$ in the phase diagram (Figure 6) is where current approval is so low that the regime will not be able to gain any approval at the next period ($f(A_i) = 0$). If A_i is a bit larger but still small, this will imply a small K_{i+1} which will lead some types on the far right to obey the regime. Meanwhile, all types on the left side of the regime (and some on the right) speak their minds, as can be seen in Figure 5. This means that an increase of A_i adds people obeying the regime only on the right side of it. However, at some point, when A_i has increased sufficiently, Δ_{i+1} will be sufficiently small so that regime followers will be added also on the left. At this point A_{i+1} becomes steeper – this is the kink – as an increase in A_i from that moment on adds obedience on both sides of the regime. The reason why $A_{i+1}(1) < 1$ is that, for any finite K_i, there will always exist sufficiently moderate types who will choose to speak their minds.[17] Put together, we get that the phase diagram has the shape depicted in Figure 6.

For any approval level A_i, an increase in the regime's force ($\bar{K}$) raises $A_{i+1} = f(A_i)$. For sufficiently small $\bar{K}$, no intersection with the 45-degree line exists, but for larger $\bar{K}$, $A_{i+1} = f(A_i)$ intersects the 45-degree line either twice or four times (bar tangency points), with the leftmost intersection being unstable, the next being stable and so on. The possibility of four intersections is precisely because of the kink.

[17] To see why this is the case, note that, when $\alpha < \beta$, $D = (x)^{\alpha}$ is steeper than $P = K(x)^{\beta}$ for sufficiently small values of x and any finite K, which implies that for a type sufficiently close to the regime it is more costly to deviate from her bliss point than to speak her mind and bear the sanctioning for doing so.

The intuitive reason for why a biased regime can employ less force yet remain stable (part 1 of the proposition) is that in the case of $\alpha < \beta \leq 1$ the regime induces silence by those with private opinions sufficiently far from it. Hence, a biased regime, whose policy is far from and thus induces silence by many in society, will have more followers than a central regime with the same force. In Figure 6 this means that biasness shifts the graph upwards.[18] This shift implies that the stable steady states move rightward while the unstable steady states move leftward in the phase diagram, implying that shocks to the approval can be larger without initiating a revolution – a biased regime is more stable.

If a regime implements popular policies, thus decreasing its bias with respect to the preferences of the population, it lowers $f(A_i)$ in the phase diagram. This may imply that a previously stable steady state disappears and a revolution is initiated (part 2 of the proposition). The reason for this is that the ones who largely agree with the regime are the ones dissenting against it by posing mild critique. Roughly speaking, when the regime implements popular policies it aligns with the views of more people thus inducing more people to speak their minds. This increases the number of individuals criticizing the regime, which weakens the regime and may ultimately be the catalytic event that starts a revolution. A similar process may be triggered by a temporary shock to the regime's force. Alternatively, a shock to the approval may lead the political system to a zone with convergence downwards. The model predicts that if the regime reacts to these events by implementing even more popular policies, as regimes under threat often naturally do, this will exacerbate its predicament. This is quite surprising but, again, stems from the fact that a popular policy induces more people to speak their minds. It further implies that implementation of *unpopular* policies could help the regime stop the revolution.

The fact that the revolution is initially driven by moderates and that the momentum of the revolution is driven by new recruits has important implications for the fragility of the revolutionary process at different stages. To see this, consider a left-biased regime which starts at the rightmost stable steady state in Figure 6. Suppose now that the force of the regime $(\bar{K})$ decreases for some reason so that the function $f(A_i)$ falls and the rightmost steady state disappears. During this phase the revolution recruits new individuals on both sides of the regime, implying a strong momentum. However, eventually no new recruits can be added from the left, as all leftists already speak their minds. This is where the dynamics reach the kink in the phase diagram, after which the momentum decreases since the new recruits come from one side only. At this point the revolution may fail if there still exists an intermediate stable steady state (i.e., if the middle region of the curve intersects the 45-degree line). In other words, the regime holds on but with less approval than it previously had (as in the leftmost A_{ss} in Figure 6). Hence, unlike the revolution described in the previous section, now the revolution is most fragile toward the end, when its momentum is

[18] More precisely, an increase in bias shifts the concave part to the left of the kink upwards and at the same time widens this part outwards in both directions.

dependent on recruits from one side only.

The revolutionary dynamics following a shift in public sentiments are illustrated in Figure 7. As mentioned earlier, this is equivalent to a shift in the regime's policy. The left-hand side (Case 1) illustrates that if the private sentiments in society shift right and the regime is left biased then this will only strengthen the regime. On the other hand, if the population's opinions shift to the left, as illustrated in Case 2, a revolution will commence. The first thing that happens is an increase in the number of moderate dissenters on the left and on the right – the horizontal bar of those speaking their minds is widened. This will weaken the regime's strength thus making it possible also for less moderate people to speak up against the regime, which increases dissent. However, since the regime is left biased, at a certain point new recruits will appear only on the right. This way, what started as a leftist revolution, following a leftward movement of public sentiments, ends up being a rightist revolution, where the center of expressed opinions is eventually to the right of the regime that collapsed and it is revealed that society was all along more rightist than the regime.

4.2 Historic examples

The revolutionary pattern just described provides a theoretically consistent explanation for an important class of revolutions and mass protests that were previously unexplained by formal theory. For example, it aligns with many of the protest movements that led to the collapses of the communist regimes in Eastern Europe in 1989-91 and to the recent Arab-Spring revolution in Egypt.

In Eastern Europe, the initial protesters in many countries were not very extreme. For instance, Hungarian communist-party leader Karoly Grosz stated that "the party was shattered not by its opponent but – paradoxically – from within" (Przeworski 1991:56). Furthermore, in Poland and Hungary, moderate dissidents instigated liberal reforms and made demands for free elections (Pfaff, 2006). In the USSR, it was the implementation of popular policies (Perestroika, to be discussed shortly) by Gorbachev that sparked the revolution. The first to protest was indeed a party insider – Boris Yeltsin – who at various meetings in 1986-1988 openly critized Gorbachev and his government for the reforms not being sufficiently far reaching (Breslauer, 2002 p.130-132). News of this insubordination spread and over these years other factions of society joined Yeltsin in criticizing and protesting against the government. The dissent against Gorbachev was in fact two-sided early on (Sanderson, 2015, p.126), as parts of the communist party disliked Gorbachev's liberalization reforms and later even tried to overturn them by staging a coup.

The trigger of the revolution in the USSR (which then spread to Eastern Europe) was the movement of the regime's policy in the direction of the liberal sentiments in society. Perestroika (i.e., economic reforms) is the equivalent of a decrease of policy bias ($|R|$) in our model. Gorbachev implemented Perestroika in the hope to revitalize and modernize the Soviet Union and thus to increase the regime's approval (Gorbachev, 1987). However, these

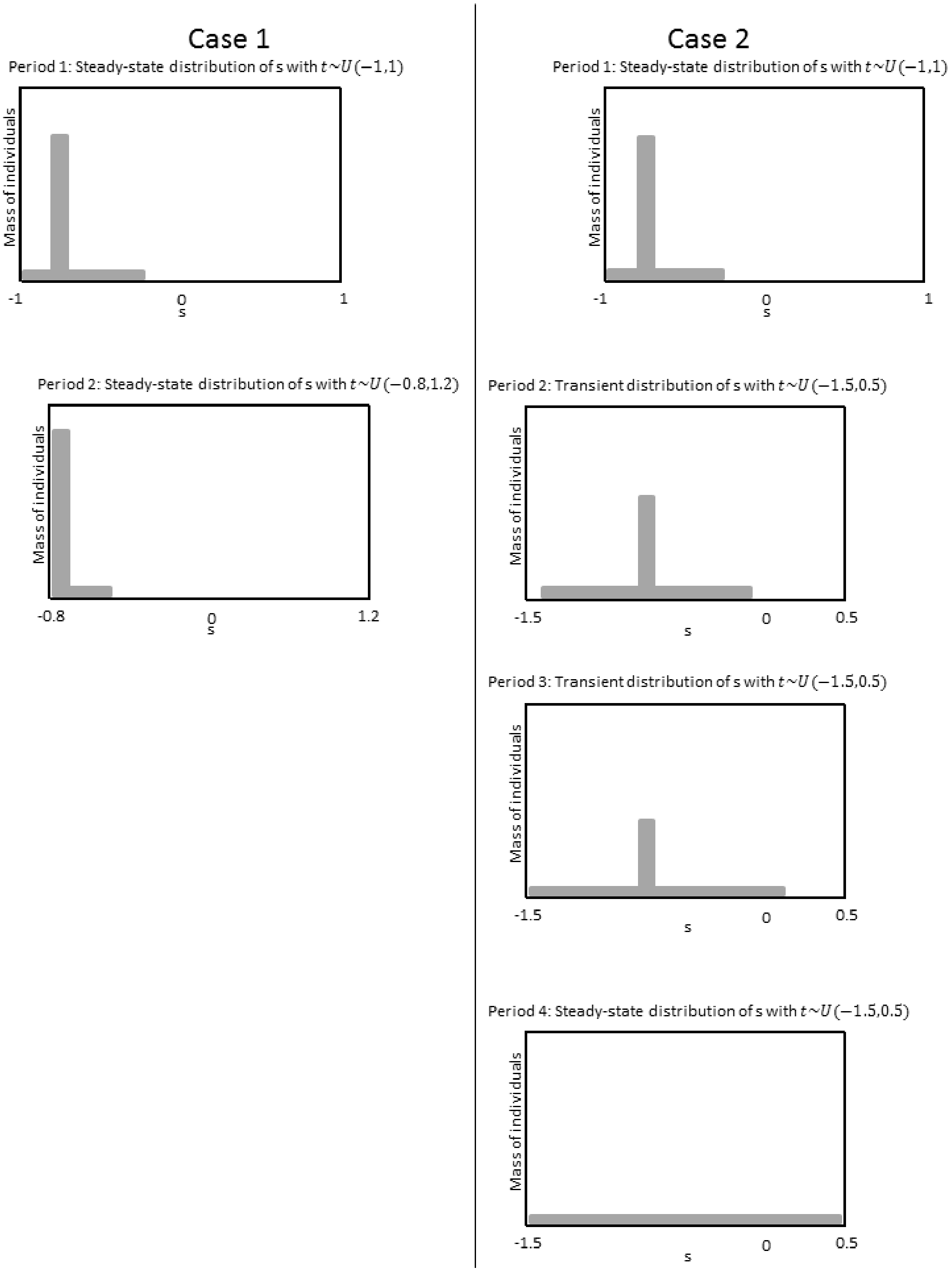


Figure 7: Distribution of stances over time when $\alpha < \beta \leq 1$. In both cases $R = -0.8$ and fixed while the distribution of types changes. In the left case, the shift of private sentiments to the left does not trigger a revolution. In the right case, the shift of private sentiments to the right does trigger a revolution starting with moderates dissenting moderately.

reforms instead were the trigger of a revolution as they unleashed social forces that brought about the dissolution of the USSR (Brown, 1997). These consequences were unintended by the leadership (see Gorbachev, 1987 p.17), came to the surprise of most experts (as documented by Kuran 1991, Lipset and Bence 1994) and are indeed counter to the predictions of the standard models of revolutions. How come popular reforms such as Perestroika can trigger a revolution? Our model provides a possible answer to this largely unresolved question. In the case described in this section, implementation of popular policies will lead to increased dissent by moderates since it becomes easier for them to speak their minds (just as Yeltsin found it easier following the Perestroika to more freely express his critique). This then paves the way for more extreme types and for more participation in dissenting. In parallel, Gorbachev implemented Glasnost (increased openness and freedom of speech). In our model this is equivalent to a decrease in $\bar{K}$, which fosters more dissent. Thus, both these reforms have the effect of undermining the regime in our model. Our answer to Przeworski's (1991, p.1) challenge to "identify the theoretical assumptions that prevented us from anticipating these developments" is thus that previous models indirectly assume that dissent is triggered only by great misalignments of preferences hence extremists have to be first to speak out and from this follows that only unpopular policies can instigate a revolution. If instead one looks at a microfounded model of dissent, like ours, it follows that also moderates may start a revolution (as observed) and hence that popular policies may be the trigger.

In Egypt in 2011, the initial protesters on the Tahrir Square were moderate liberals and moderate conservatives (Al Jazeera 2011, Lesch 2011). The most extreme factions (i.e., the Muslim Brotherhood and the Salafi movement) were not present in the protests initially. They only joined later.[19] Beyond the progress from moderate to extreme, another important feature of this class of revolutions is that the undermining of the regime is initiated by individuals with moderate views from *both* sides of the political spectrum (unless the regime is so biased that on one of its sides there are no individuals). This implies that regimes may be undermined by truly "strange bedfellows", in the sense that they are pulling the public opinion in two different directions. This was a clear pattern in the Arab-Spring revolution in Egypt. The protesters on the Tahrir Square consisted of some who suggested that Mubarak was not sufficiently liberal and of others who said he was not sufficiently conservative and religious. While the spark may have been a shift in private opinions towards more liberalism (a leftward movement of the opinion axis when moving from the first to the second schedule of case 2 in Figure 7), the later elections showed that in fact Egyptian society as a whole was even more conservative than Mubarak's regime (in line with the description in Case 2 in Figure 7, where the average opinion after the shift is to

[19] For instance, a BBC news profile on the Muslim Brotherhood reports that initially "(t)he group's traditional slogans were not seen in Cairo's Tahrir Square. But as the protests grew and the government began to offer concessions, including a promise by Mr Mubarak not to seek re-election in September 2011, Egypt's largest opposition force took a more assertive role" (BBC, 2013).

the right of $R = -0.8$, which represents Mubarak's regime in that figure). This way, as predicted by the model, what started as mainly a leftist (liberal) revolution ended up being a rightist (conservative) revolution instead.

5 Extremists starting with moderate dissent

5.1 Analysis

The final case is when $\alpha > 1$ and $\beta \geq 1$.[20] This case shares the moderate-to-extreme progress of public statements during the revolution with the case of the previous section, while sharing the leading role of extremists in the revolution with the class of revolutions described in Section 3.

An important feature of this case is that $\beta > 1$, which represents a regime that is tolerant to small dissent while punishing harshly larger dissent. By differentiating (4) twice it is immediate that the second-order condition holds so that each type has an inner solution.[21] This means that each type compromises between fully obeying the regime and speaking her mind. This is intuitive since when the regime is tolerant toward small dissent, the citizens do not have an incentive to keep completely silent. At the same time, when D is convex, the citizens are lax about small deviations from their bliss points and hence do not mind compromising a little. Furthermore, extremists dissent more than moderates since the convexity of D makes large deviations from one's bliss point very costly.

The result that extremists are compromising yet dissent more than the moderates has important implications for the stability of regimes and for the revolutionary dynamics as expressed in the following proposition.

Proposition 4 *When $\alpha > 1$, $\beta \geq 1$:*

1. ***Existence of a stable steady state:*** *A stable regime exists iff it employs sufficient force, and the more biased its policy is the more force it needs to employ.*
2. ***Catalytic events:*** *A revolution may start following a shock to the regime's approval or force or following implementation of unpopular policies.*
3. ***Dynamics of participation:***
   (a) *All types participate at all time periods during a revolution (unless $\beta = 1$, in which case only the most extreme types participate initially).*
   (b) *For any regime with $|R| \neq 1$ the revolution will be two-sided throughout.*

[20] When α equals exactly one, some small technicalities need to be kept in mind. The results with respect to what will be presented are however the same, so we will simply ignore this case in our analysis.

[21] In case $\beta = 1$ types close to the regime fully obey it while types far have an inner solution.

4. ***Dynamics of statements**: The most extreme types are the ones dissenting the most at all time periods. They start moderately and increase their dissent over time.*

Again, we start by explaining the dynamics of the revolutions (parts 3 and 4 of the proposition) since this largely determines what makes a regime stable and which events may initiate a revolution. Part 4 of the proposition says that, once the revolution starts, dissent intensifies over time, as depicted for a stylized example in Figure 8. The intuition for this is that a convex punishment ($\beta > 1$) implies a relatively heavy sanctioning on extreme dissent. Hence, very extreme dissent will be absent initially. During the process of a revolution, as the approval and hence also the strength of the regime fall, all types are induced to dissent more, and in particular it becomes possible to express views that are more extreme than was possible before. This further weakens the regime, causing more dissent and so on. As expressed in part 3, the ones who are dissenting the most are the extremists. This is an important difference between this class of revolutions and the revolution starting with moderates described in Section 4. In contrast to the revolution starting with moderates, in which the most extreme types remain silent for a very long time, here the extremists are the ones constantly pushing the freedom of speech, backed-up from behind by the moderates.

The dynamics of approval are depicted for a stylized example with $\beta > 1$ in Figure 9. As in the previous phase diagrams, the intertemporal-dynamics function $A_{i+1} = f(A_i)$ is first flat at zero (unless $R = 0$) and then increases. As can be seen in the figure, $A_{i+1}(1) < 1$, reflecting that there cannot be full obedience when $\beta > 1$.[22] Depending on the values of $\bar{K}$ and R, the intertemporal-dynamics function $f(A_i)$ has either zero or two intersections with the 45-degree line (bar tangency points), but never more than two. Considering the case of two intersections as in the figure, the fact that $A_{i+1}(1) < 1$ implies that the right intersection is stable while the left is unstable and the meeting point at zero is stable too.

It is intuitive that more force (i.e., larger $\bar{K}$) increases the public approval (A_{i+1}) of the regime since it decreases dissent for any level of previous approval (A_i). This shifts the function $f(A_i)$ up in the phase diagram, thus enabling the existence of a stable steady state. An increase in the bias of the regime's policy, on the other hand, decreases approval. To see why, compare two cases, one where the regime is at $R = 0$ and one where it is at $R = -1$. The former has a mass of types on its left with private opinions at distances between zero and one and similarly on the right. Switching from $R = 0$ to $R = -1$ is like replacing the mass of types on the left with a mass of types on the right, but now with private opinions at distances of one to two. That is, we replace individuals who are moderately critical to the regime in private with individuals who are much more critical. Since types who are more extreme dissent more, we get that approval is decreasing in the bias of R. Hence, biasness shifts the $f(A_i)$ function down, implying that larger bias of the regime has to be compensated for by the employment of more force for a stable steady state to exist (part 1). Furthermore, by lowering the approval function, biasness reduces the distance between the

[22] If β exactly equals 1 there can be full obedience, provided that $\bar{K}$ is sufficiently large.

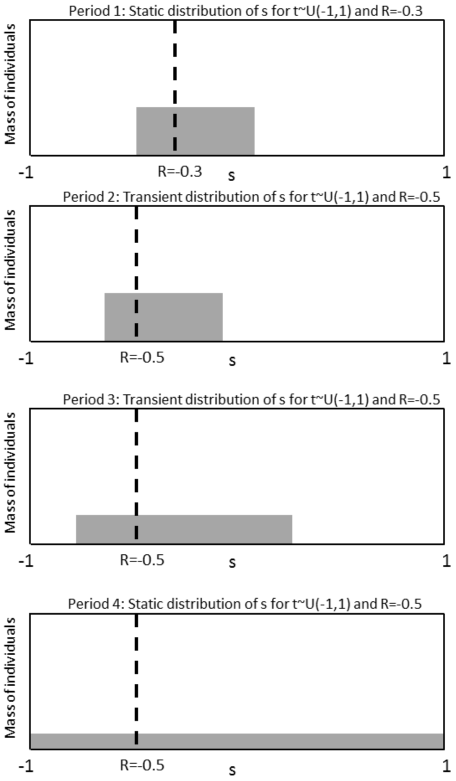


Figure 8: Distribution of stances over time in a stylized case of a revolution starting with extremists dissenting moderately ($\alpha > 1$, $\beta \geq 1$). The regime starts at $R = -0.3$ and after the first period moves to $R = -0.5$ (which triggers the revolution) and stays there, while the distribution of types is constant. The diagram depicts the case of $\alpha = \beta$ for ease of exposition

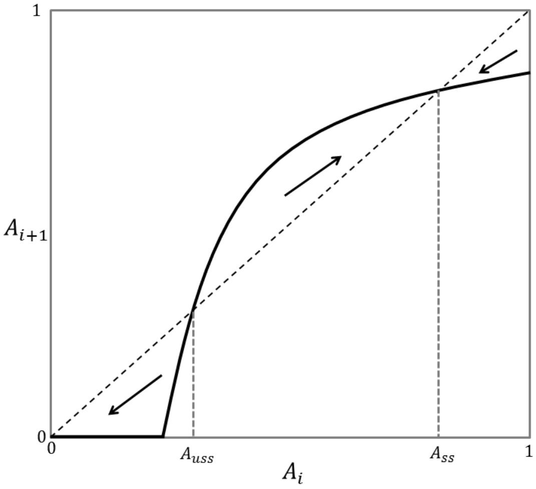


Figure 9: Stylized phase diagram for the case $\alpha > 1$ and $\beta \geq 1$ when the regime is biased. The full line depicts the intertemporal-dynamics function $A_{i+1} = f(A_i)$ and the dashed line depicts 45-degree line where $A_{i+1} = A_i$. The vertical lines depict the stable (A_{ss}) and unstable (A_{uss}) steady states.

unstable and stable steady states, hence increases the vulnerability of the regime to shocks to its approval or force – biased regimes are less stable.

It is worth noting that, if a revolution starts, the regime will always eventually collapse (because there is no other stable steady state to the left of the initial state besides the one at $A = 0$).[23] Unlike the two previous classes of revolutions, here the revolution never loses its momentum since it is the gradual shift of statements that drives it instead of recruitment of new protesters. Hence, once a revolution has started it will always succeed, unless the regime reacts on time by either increasing its force (e.g., by recruiting more troops) or implementing popular policies to pease the population.

5.2 Historic examples

The main feature that distinguishes this class of revolutions is that here during the whole revolution it is that extremist dissenters are the ones expressing the strongest criticism, while constantly increasing their dissent against the regime and pulling other factions of society along with it.

As a historic example, consider the April Revolution in South Korea, 1960. We provide here a summary of the gradual escalation of events in this student-led revolution (taken from Kim, 1996) from the minor protest against a governmental intervention to the mass demand for resignation of the president. The protest movement began on February 28th 1960, two

[23] $R = 0$ with $\alpha > 2$ is a special case where there is no stable steady state at 0 because $f(A_i)$ starts with an infinite slope. In this case, $f(A_i)$ is concave throughout with exactly one stable steady state with $A \in]0, 1[$, corresponding to a regime that cannot collapse.

weeks before the planned (rigged) national elections, when students from Taegu marched into the streets in protest against a governmental directive to attend school on a Sunday. On election day, March 15, angry citizens of the city of Masan, whose names had been removed from the voter registration roster, marched into the city hall asking that voting slips be given to them. About a month after the elections, on 18-19th of April, students gathered in front of the principal government buildings with a more far-reaching demand for new, fair elections. A week later, dissent further increased to loudly demanding the resignation of the president. In the evening dissent escalated further, when people rioted in an attempt to overthrow the regime. President Rhee indeed resigned the following morning. This revolution seems to have been triggered by implementation of unpopular policies as President Rhee was trying to become an autocrat against the will of the people.

A similar chain of events characterized the student-led protests on Tiananmen Square in Beijing in 1989 (for a detailed account see Zhao 2001), with the important difference being that in the case of Beijing the protests did not develop into a successful revolution that would end with a change of the regime. The former politician, Hu Yaobang, who was popular among students, passed away on April 15 1989 and this led a large number of students to mourn his death on that day (Pan, 2008). Two days later, a commemoration (which was considered more dissenting than individual mourning) was organized. This organization quickly evolved into a declaration of demands for political reform and thereafter, on April 18, to a sit-in where students demanded to meet with the leadership of the political party. On April 21, students began organizing themselves formally into unions and some workers into a federation, writing texts challenging the regime (Walder and Xiaoxia, 1993) and the next day serious rioting broke out in several places. Five days later, the Autonomous Student Union staged a march to the square breaking through police lines after which the leaders of the union, Wang Dan and Wu'erkaixi, called for more radical measures to regain momentum. This led to hunger strikes and also to the expressed support for the strikes by others who did not themselves strike. On May 17–18, around a million Beijing residents, including low-ranked representatives of the regime such as party officials and police officers, demonstrated in solidarity of the hunger strikers. This was a sign of the decreased approval of the regime, as predicted by our model, with the sign of the regime's consequent weakening being the increasingly open and positive reports about the protests in the media. All this time, the soft approach of the regime – of showing sympathy toward the demonstrators and looking for a dialogue with them – as advocated by Zhao Ziyang, the General Secretary of the Communist Party, was giving the tone. This approach of containment of the moderately deviant expressions is predicted, if not interrupted, to have led to the eventual collapse of the regime. However, on May 17, a leadership meeting was called, where Zhao Ziyang's concessions-based strategy was thoroughly criticized and it was decided to declare martial law. In terms of our model, this was a decision to substantially increases $\bar{K}$. The implementation of the martial law that led to the heavy-handed crackdown on the protests on June 4 eventually stopped the mass protests.[24]

6 Further results and empirical predictions

This far, we have provided examples of actual revolutions and mass protests that fit the different patterns resulting from our model. However, we have not provided evidence that the parameter combinations necessary for producing each of the patterns is fulfilled. The simple reason for this is that the focus of our paper is theoretical and that attaining the parameter values requires in-depth study of sufficiently many revolutions. Furthermore, since many other factors may affect the shape of revolutions, showing that the parameters do fit a specific case cannot be considered sufficient evidence that the model is correct neither would a lack of fit of parameters for a single case be evidence that it is not correct – to test our model one needs to perform an empirical test with many observations. In this section we discuss some results that apply generally to the whole model and formulate them in the form of empirical predictions. These predictions relate mainly to β and α.

Starting with β, the curvature of the sanctioning system, the previous propositions show that it is only for sufficiently small β that the revolution starts with expression of extreme views and continues with expression of more moderate views.

- **Prediction 1**: Holding all else fixed, the less a regime distinguishes between large and small dissent (i.e., the lower is β), the more likely it is that the dynamics of the revolution will be such that the most dissenting expressions appear right from the start.

This prediction holds also if α is heterogeneous between individuals – the smaller is β, the more likely it is that it is smaller than any given α and hence, conditioned on $\beta \leq 1$, that it will induce extremists to be the first speaking their minds against the regime. In principle, the sanctioning structure of the regime is observable. One way of obtaining a proxy for β is to look at past protests and how the regime sanctioned deviant expressions: if only extreme dissent was punished then β is large, but if the regime punished all dissent, large and small, roughly the same, then β is small. One can then look at how the current revolution evolves.

[24] In the case of the protesters on Tiananmen Square it is hard to say what was the catalytic event starting the protests. The regime had been implementing popular policies throughout the 1980s. These reforms were initiated by Chairman Deng Xiaoping, the successor of Mao Zedong, and were indeed generally well received by the public. According to our model, such policies would make a revolution *less likely* and hence cannot be the cause of the revolution. We emphasize the words "less likely" since this policy change does not mean a revolution cannot happen at all. What could have led to the start of protests despite the policy change is that the private views of the population had shifted even more, that the regime had reduced its force, or that there had been a temporary shock to the approval of the regime. Perhaps the latter is the most likely explanation, whereby those who initially gathered to mourn the death of Hu Yaobang managed to overcome the collective-action problem during their meeting. This kind of shock to the approval is exogenous in our model.

As for α, the curvature of the cost of bliss-point deviations, it is harder to observe (unless D is an economic loss, whereby α in principle can be derived from economic structures). Hence we need to identify two observable variables that depend on α and that the model predicts should be related in a certain way. One such prediction relates to policy changes. According to the previous propositions, implementation of popular policies should trigger a revolution if and only if $\alpha \leq \{1, \beta\}$, in which case the revolution will start with moderate forces that will gradually recruit the extremists. This leads to the following prediction.

- **Prediction 2:** Holding all else fixed, there is a positive correlation between implementation of popular policies and revolutions that are initiated by moderates (or regime insiders).

In principle it is observable whether a policy change is in the interest of most of the population or not. It is also observable whether a revolution starts afterwards and, if so, the prediction says that the initial protesters should be moderate (as in the USSR and Egypt).

The third prediction is a mirror image of Prediction 2 and has to do with the effectiveness of the regime's response to the revolution. A very common reaction of regimes that see an escalation in dissent and their approval deteriorating is to offer reforms – popular policies that are meant to pease the population and consequently cool down the angry citizens. Our model predicts however that this measure will not be effective against the second class of revolutions. For instance, if Gorbachev would have implemented even more popular policies, this would not have helped him stay in power.

- **Prediction 3:** Revolutions that start with moderates will not subside following implementation of popular policies.

Finally, a general point – about the timing of regime reactions, martial law and rule by decree – is revealed when looking at the phase diagrams in Figures 3, 6 and 9. To illustrate it, suppose there is a shock to the regime's force $(\bar{K})$ which lowers the function $A_{i+1} = f(A_i)$ sufficiently so that the old stable steady state disappears (a revolution is initiated). Suppose further that the regime intends to use increased force to cool down the protests. Then, if a few time periods have already passed since the revolution started, it may not be sufficient to simply restore the old level of force in order to achieve upward convergence of approval. This is since the approval may have deteriorated to a level A_i where also the old intertemporal-dynamics function implies downward convergence. The model thus predicts that what the regime has to do in this case is to overshoot its force, as is indeed often done by the implementation of martial law or rule by decree. What is further interesting is that it is sufficient that these extraordinary measures be temporary – once the approval has grown sufficiently, the martial law can be abandoned without risking a resurrection of the revolution. This explains why harsh temporary measures often tend to work, like is illustrated by the governmental intervention in Tiananmen Square.

The same logic applies to other measures that raise the approval function, such as the implementation of new policies (unpopular ones in the second class of revolutions and popular ones otherwise) – the regime has to temporarily overshoot with these policies but can then restore the old policies once the protests have calmed down. Moreover, the timing of the intervention of the regime is crucial, as expressed by the following prediction.

- **Prediction 4:** Holding all else fixed, the later the regime implements a given increase in force, the more likely the revolution is to succeed.

The prediction follows directly from the phase diagrams and stems from the fact that the longer the regime waits with reacting to a revolution or protests, the lower the approval level (A_i) will be at the time of reaction, and hence a given increase of $\bar{K}$ is less likely to ensure that the approval function $A_{i+1} = f(A_i)$ rises sufficiently to imply upward convergence.[25]

7 Conclusions

This paper presents a unified framework to explain three classes of popular revolutions and mass protests that have been observed historically: 1) revolutions starting with extremists dissenting extremely; 2) revolutions starting with moderates dissenting moderately; and 3) revolutions starting with extremists dissenting moderately. Earlier models invariably predict that the revolution will be initiated by extremists and are silent about the extent to which each individual will dissent and how this will change over time. Thus they cannot distinguish between the first class (Iran 1978-79) and the third class (Tiananmen Square in 1989) of revolutions and are outright inconsistent with the second class of revolutions (Egypt in 2011 and Eastern Europe in 1989-91).

The classification into three classes of revolutions is shown to be exhaustive in our model. It spans the parameter space of α and β, the parameters that capture the curvature of the two costs affecting the individual choice of dissent during a revolution: the cost of deviating from her privately held opinion (or economic interest) and the cost of dissenting against the regime. Each class of revolutions has its own unique set of attributes, characterizing who in society – moderates or extremists – initiates the revolution, how it progresses to other parts of society, which views are expressed by participants at various stages, how the regime may unknowingly trigger the revolution and what it can or cannot do to stop the revolution at different stages. The overarching pattern is that the curvature of the regime's punishment affects how extreme the initial dissent in a revolution will be; and the curvature of the cost of deviating from an individual's privately held opinion affects which types – extremists or moderates – will start the revolution.

It would be presumptuous on our side to actually claim that all real-life revolutions follow one of the three patterns. Our analysis abstracts from important real-life factors

[25] This result is not specific to our model but can be obtained by other models, e.g., Olsson-Yaouzis (2012).

such as intervention of outside forces; conflicts within different revolutionary groups about the targets of the revolution; changes in the regime's leadership during the revolution or endogenous reactions of the regime (as was shown to be crucial in the case of the protests on Tiananmen Square); and heterogeneity in the private costs (α). However, as the historic examples provided in the paper demonstrate, our model *is* able to capture many important aspects of real revolutions that cannot be captured with the existing models and accordingly we provide empirical predictions on the progress of revolutions, the catalytic events leading to them and the effective and ineffective responses of regimes. In particular, these predictions do not require a homogenous α in society. Naturally, our framework can be used to study other aspects of revolutions and mass protests, where further questions could be answered and more parameters could be endogenized.

A Analytical derivations and proofs

A.1 Some auxiliary results

Using equation (6) with $\lambda = 1$ we have

$$\begin{aligned} A_{i+1} &= \begin{cases} 1 - \Psi\left(s_i; R, A_i\right) \text{ when } 1 - \Psi\left(s_i; R, A_i\right) \geq 0 \\ 0 \text{ when } 1 - \Psi\left(s_i; R, A_i\right) < 0 \end{cases} \text{where} \quad (7) \\ \Psi\left(s_i; R, A_i\right) &\equiv \int_{-1}^{1} \left|s_i^*\left(t\right) - R\right| dt. \quad (8) \end{aligned}$$

A.2 Individual stances

The individual minimizes the loss function given by (4), (1) and (3) when $K > 0$. Using the implicit function theorem we get the following derivatives of $s^*(t)$ in inner solutions:

$$\frac{ds^*}{dt} = \frac{D''\left(t - s^*\right)}{P''\left(s^*\right) + D''\left(t - s^*\right)} \quad (9)$$

Let t_l and t_h denote the left and the right edges of distribution of types, and let

$$\Delta \equiv K^{\frac{1}{\alpha - \beta}}.$$

A.2.1 Case (1): $\max\left\{\alpha, \beta\right\} \leq 1$

The second-order condition of the loss function is positive when $\alpha < \beta \leq 1$ or $\beta < \alpha \leq 1$, which implies that any inner extreme point is a maximum. The corner solutions are then either $L\left(s = R\right) = \left|t - R\right|^{\alpha}$ or $L\left(s = t\right) = K\left|t - R\right|^{\beta}$. When $\beta < \alpha$ this implies that $L\left(s = R\right) < L\left(s = t\right)$ iff $\left|t - R\right| < \Delta$, and so $s^*\left(t\right) = t$ iff $\left|t - R\right| \geq \Delta$, and $s^*\left(t\right) = R$ iff $\left|t - R\right| < \Delta$. When $\alpha < \beta$ the converse holds,[26] with $s^*\left(t\right) = t$ iff $\left|t - R\right| \leq \Delta$, and $s^*\left(t\right) = R$ iff $\left|t - R\right| > \Delta$.

[26] Since then $\frac{1}{\alpha - \beta} < 0$, hence when solving for $K^{\frac{1}{\alpha - \beta}}$ the inequality flips direction.

A.2.2 Case (2): $\beta < 1 < \alpha$

We perform the proof for $t \geq R$. The opposite case is similar. We will prove that if $t_h - t_l > 2\Delta$, then types close enough to the regime fully conform, while types far from the regime choose an inner solution and $|s^*(t) - R|$ is increasing for them. Along the way we will also show that for a sufficiently narrow range of types, the distribution is degenerate at R.

We will first show that the only relevant corner solution is $s^* = R$. In order to find the global minimum for a type t, we first need to investigate the behavior of $L(s,t)$ at $s = t$ and $s = R$.

$$L'(s,t) = -\alpha (t-s)^{\alpha-1} + \beta K (s-R)^{\beta-1}$$

Hence $\lim_{s\to R} L'(s,t) = \infty$ and $L'(t,t) = \beta K (t-R)^{\beta-1} > 0$. Therefore $s = R$ may be a solution to the minimization problem while $s = t$ is not. The candidate solution $s = R$ will now be compared to potential local minima in the range $]R,t[$. In inner solutions $L'(s,t) = 0$ and hence we get

$$\begin{aligned} \alpha (t-s)^{\alpha-1} &= \beta K (s-R)^{\beta-1} \\ &\Rightarrow (t-s)^{\alpha-1} (s-R)^{1-\beta} = \beta K/\alpha. \end{aligned} \tag{10}$$

Define

$$\Phi(s) \equiv (t-s)^{\alpha-1} (s-R)^{1-\beta}.$$

For the existence of an inner min point for a given t it is necessary that $\Phi(s) = \beta K/\alpha$ for some $s \in]R,t[$. Note that as $t \to R$ both $(t-s)^{\alpha-1}$ and $(s-R)^{1-\beta}$ approach zero implying $\Phi(s) < \beta K/\alpha$ for all $s \in]R,t[$. Hence types with sufficiently small $|t-R|$ do not have an inner local min point and they choose $s^* = R$.

For sufficiently large $|t-R|$ it may be that $\Phi(s) = \beta K/\alpha$ for some $s \in]R,t[$ which we investigate next. Note that, for given t, $\Phi(s)$ is strictly positive in $]R,t[$, and that $\Phi(s,t) = 0$ at both edges of the range (i.e. at $s = R$ and at $s = t$). This means that $\Phi(s)$ has at least one local maximum in $]R,t[$. We now proceed to check whether this local maximum is unique:

$$\Phi'(s) = (t-s)^{\alpha-2} (s-R)^{-\beta} [(1-\beta)(t-s) - (\alpha-1)(s-R)]$$

Since $(t-s)^{\alpha-2} (s-R)^{-\beta}$ is strictly positive in $]R,t[$, and $[(1-\beta)(t-s) - (\alpha-1)(s-R)]$ is linear in s, positive at $s = R$ and negative at $s = t$, $\Phi'(s) = 0$ exactly at one point at this range (i.e. a unique local maximum of $\Phi(s)$ in $]R,t[$). From the continuity of $\Phi(s)$ we get that if the value of $\Phi(s)$ at this local maximum is greater than $\beta K/\alpha$, then $L(s,t)$ has exactly two extrema in the range $]R,t[$. From the positive values of $L'(s,t)$ at the edges of this range we finally conclude that the first extremum (where $\Phi(s)$ is rising) is a maximum point of $L(s,t)$, and the second extremum (where $\Phi(s)$ is falling) is a minimum point of $L(s,t)$. The global minimum of $L(s,t)$ is therefore either this local minimum (i.e. an inner solution), or $s = R$ (i.e. a corner solution). If however the value of $\Phi(s)$ at its local maximum point is smaller than $\beta K/\alpha$, then there is no local extremum to $L(s,t)$ in the range $]R,t[$, and therefore $s = R$ is the solution to the minimization problem.

Next we show that if $t_h - t_l > 2\Delta$ then there exists a type who is far enough from the regime to choose the inner solution. First, note that the distance from the regime to the type who is the most remote from it is larger than Δ when $t_h - t_l > 2\Delta$. Suppose this type is $t_{h.}$. Then, comparing only the two corner solutions this type can choose, we get

$$L(R, t_h) - L(t_h, t_h) = |t_h - R|^{\alpha} - K\, |t_h - R|^{\beta},$$

which is strictly positive when $|t_h - R| > \Delta = K^{\frac{1}{\alpha-\beta}}$ and $\beta < \alpha$. This implies that t_h does not choose the corner solution of R, hence must choose an inner solution.

Now we show that if there exists any type t_0 who chooses the inner solution then all types with $t > t_0$ have an inner solution. We also show that types close enough to the regime fully conform, and that in the range of inner solutions $|s^*(t) - R|$ is increasing in t. First note that $\Phi(s)$ is increasing in t, so if there exists a local minimum of $L(s, t_0)$ for some t_0, then there exists a local minimum of $L(s, t)$ for $t > t_0$ too. Also note that for all $s \in]R, t[$ $\Phi(s)$ is increasing in t and that $\lim_{t\to\infty} \Phi(s, t) = \infty > \beta K/\alpha$, implying an inner local min point exists for a broad enough range of types. Second, if there is an inner solution to the minimization problem for some t_0 then there is also an inner solution to the minimization problem for $t > t_0$. To see this let $\Delta L \equiv L(R, t) - L(\tilde{s}, t)$, where $\tilde{s}$ is the stance at which $L(s, t)$ gets the local minimum. Type t prefers the inner solution to the corner solution if and only if ΔL is positive. Thus we need to show that ΔL is increasing in t and so if ΔL is positive for t_0 then it is positive for $t_1 > t_0$ too.

$$\Delta L = (t - R)^{\alpha} - (t - \tilde{s})^{\alpha} + K (\tilde{s} - R)^{\beta}$$

Differentiating ΔL with respect to t yields

$$\Delta L'_t = \alpha (t - R)^{\alpha-1} - \left[\alpha (t - \tilde{s})^{\alpha-1} \left(1 - \frac{d\tilde{s}}{dt}\right) + \beta K (\tilde{s} - R)^{\beta-1} \frac{d\tilde{s}}{dt}\right].$$

Using the first-order condition (10)

$$\begin{aligned}\Delta L'_t &= \alpha (t - R)^{\alpha-1} - \left[\alpha (t - \tilde{s})^{\alpha-1} \left(1 - \frac{d\tilde{s}}{dt}\right) + \alpha (t - \tilde{s})^{\alpha-1} \frac{d\tilde{s}}{dt}\right] \\ &= \alpha (t - R)^{\alpha-1} - \alpha (t - \tilde{s})^{\alpha-1} > 0\end{aligned}$$

Differentiating once more

$$\Delta L''_t = \alpha (\alpha - 1) \left[(t - R)^{\alpha-2} - (1 - d\tilde{s}/dt)(t - \tilde{s})^{\alpha-2}\right].$$

By equation (9) we have that $\frac{d\tilde{s}}{dt} > 1$ in an inner solution when P is concave, and so $\Delta L''_t > 0$. Hence ΔL is strictly increasing and strictly convex, implying that for a broad enough range of types (in particular larger than 2Δ, as shown above), types sufficiently far from the regime have an inner solution where $\frac{ds^*}{dt} > 1$, and so $|s^*(t) - R|$ is increasing in t at the range of inner solutions.

A.2.3 Case (3): $\alpha < 1 < \beta$

We perform the analysis for $t \geq R$. The opposite case is similar. We will first show that the only relevant corner solution is $s^* = t$, then that types close to the regime choose this corner solution. In order to find the global minimum we first need to investigate the behavior of $L\left(s,t\right)$ near the corner solutions.

$$L'\left(s,t\right) = -\alpha\left(t-s\right)^{\alpha-1} + \beta K\left(s-R\right)^{\beta-1}$$

Hence $L'\left(R,t\right) < 0$ and $L'\left(t,t\right) < 0$ since $\alpha < 1$. Therefore $s = t$ may be a solution to the minimization problem while $s = R$ is not. The candidate solution $s = t$ will now be compared to potential local minima in the range $[R,t]$. In inner solutions $L'\left(s,t\right) = 0$ and hence we get

$$\begin{aligned} \alpha\left(t-s\right)^{\alpha-1} &= \beta K\left(s-R\right)^{\beta-1} \\ &\Rightarrow \left(t-s\right)^{\alpha-1}\left(s-R\right)^{1-\beta} = K\beta/\alpha \end{aligned} \tag{11}$$

Define

$$\Phi\left(s\right) \equiv \left(t-s\right)^{\alpha-1}\left(s-R\right)^{1-\beta}.$$

For the existence of an inner min point it is necessary that $\Phi\left(s\right) = \beta K/\alpha$ for some $s \in]R,t[$. Since $\alpha < 1$ and $\beta > 1$ follows that $\Phi = \beta K/\alpha$ for all s when t is sufficiently small and K is finite. Hence, sufficiently small t do not have an inner local min point which implies $s^* = t$ is the global optimum for these types. Notice that $\Phi(s)$ is strictly positive in $]R,t[$, and that $\Phi(s) \to \infty$ at both edges of the range (i.e. at $s = R$ and at $s = t$). This means that $\Phi(s)$ has at least one local minimum in $]R,t[$. We now proceed to check whether this local minimum is unique:

$$\Phi'\left(s\right) = \left(t-s\right)^{\alpha-2}\left(s-R\right)^{-\beta}\left[\,\left(1-\beta\right)\left(t-s\right) - \left(\alpha-1\right)\left(s-R\right)\,\right].$$

Since $\left(t-s\right)^{\alpha-2}\left(s-R\right)^{-\beta}$ is strictly positive in $]R,t[$, and $[\left(1-\beta\right)\left(t-s\right) - \left(\alpha-1\right)\left(s-R\right)]$ is linear in s, negative at $s = R$ and positive at $s = t$, $\Phi'(s) = 0$ exactly at one point at this range (i.e. a unique local minimum of $\Phi(s)$ in $]R,t[$).

From the continuity of $\Phi(s)$ we get that if the value of $\Phi(s)$ at this local minimum is smaller than $\beta K/\alpha$, then $L(s,t)$ has exactly two extrema in the range $]R,t[$. From the negative values of $L'(s,t)$ at the edges of this range we finally conclude that the first extremum (where $\Phi(s)$ is falling) is a minimum point of $L(s,t)$, and the second extremum (where $\Phi(s)$ is rising) is a maximum point of $L(s,t)$. The global minimum of $L(s,t)$ is therefore either this local minimum (i.e. an inner solution), or $s = t$ (i.e. a corner solution). If however the value of $\Phi(s)$ at its local minimum point is larger than $\beta K/\alpha$, then there is no local extremum to $L(s,t)$ in the range $]R,t[$, and therefore $s = t$ is the solution to the minimization problem.

Next we show that if $t_h - t_l > 2\Delta$ then there exists a type who is far enough from the regime to choose the inner solution. First, note that the distance from the regime to the type who is the most remote from it is larger than Δ. Suppose this type is $t_{h.}$. Then, comparing only the two corner solutions this type can choose, we get

$$L(R,t_h) - L(t_h,t_h) = \left|t_h - R\right|^{\alpha} - K\left|t_h - R\right|^{\beta},$$

which is strictly negative when $\left|t_h - R\right| > \Delta = K^{\frac{1}{\alpha-\beta}}$ and $\alpha < \beta$. This implies that t_h does not choose the corner solution of $t = t_h$, hence must choose an inner solution.

We now show that if there exists any type t_0 who chooses the inner solution, then all types with $t > t_0$ have an inner solution too. We also show that in the range of inner solutions $s^*(t)$ is decreasing in t. First notice that $\Phi(s)$ is decreasing in t, so if there exists a local minimum of $L(s,t_0)$ for some t_0, then there exists a local minimum of $L(s,t)$ for $t > t_0$ too. Also note that $\Phi(s)$ is decreasing in t with $\lim_{t\to\infty} \Phi(s) = 0 < \beta K/\alpha$ (for $s \in]R,t[$), implying that an inner local minimum exists for a sufficiently large t. Second, if there is an inner solution to the minimization problem for some t_0, then there is also an inner solution to the minimization problem for $t > t_0$. To see this let $\Delta L \equiv L(t,t) - L(\tilde{s},t)$, where $\tilde{s}$ is the stance at which $L(s,t)$ gets the local minimum. Type t prefers the inner solution to the corner solution if and only if ΔL is positive. Thus we need to show that ΔL is increasing in t and so if ΔL is positive for t_0 it is positive for $t > t_0$ too.

$$\Delta L= K(t-R)^{\beta} - \left[(t-\tilde{s})^{\alpha} + K(\tilde{s}-R)^{\beta}\right].$$

Differentiating ΔL with respect to t yields

$$\Delta L'_t = K\beta(t-R)^{\beta-1} - \left[\alpha(t-\tilde{s})^{\alpha-1}\left(1-\frac{d\tilde{s}}{dt}\right) + \beta K(\tilde{s}-R)^{\beta-1}\frac{d\tilde{s}}{dt}\right].$$

Using the first-order condition

$$\begin{aligned}\Delta L'_t &= K\beta(t-R)^{\beta-1} - \left[\beta K(\tilde{s}-R)^{\beta-1}\left(1-\frac{d\tilde{s}}{dt}\right) + \beta K(\tilde{s}-R)^{\beta-1}\frac{d\tilde{s}}{dt}\right] \\ &= K\beta(t-R)^{\beta-1} - \beta K(\tilde{s}-R)^{\beta-1} > 0 \text{ when } \beta > 1.\end{aligned}$$

Differentiating once more

$$\Delta L''_t = K\beta(\beta-1)\left[(t-R)^{\beta-2} - \beta K\frac{d\tilde{s}}{dt}(\tilde{s}-R)^{\beta-1}\right].$$

By equation (9) we have that $\frac{d\tilde{s}}{dt} < 0$ in an inner solution when D is concave, and so $\Delta L''_t > 0$. Hence ΔL is strictly increasing and strictly convex, implying that for a broad enough range of types, types sufficiently far from R have an inner solution. Moreover, at this subrange of types, $\frac{ds^*}{dt} < 0$ by (9) when D is concave (the denominator is positive in inner solutions by the second-order condition). This implies that $s^*(t)$ is decreasing in the subrange of types with inner solutions.

A.2.4 Case (4): $\min\{\alpha,\beta\} \geq 1$

The minimization problem of type t is symmetric around R, so we will present the first- and second-order conditions for an inner solution only for $t \geq R$.

$$\begin{aligned} -\alpha(t-s)^{\alpha-1} + \beta K(s-R)^{\beta-1} &= 0 & (12)\\ (\alpha-1)\alpha(t-s)^{\alpha-2} + (\beta-1)\beta K(s-R)^{\beta-2} &> 0 & (13)\end{aligned}$$

We perform the proof first for $\alpha,\beta > 1$, and then for the special cases of $1 = \beta < \alpha$ and $1 = \alpha < \beta$.

$\alpha,\beta > 1$: That every t has a unique inner solution can be easily verified using equations (12) and (13). Moreover, by applying the implicit function theorem to equation (12), we get that $ds^*/dt > 0$, hence $|s^*(t) - R|$ is increasing in the distance to the regime.

$1 = \beta < \alpha$: It is easy to verify that types sufficiently close to the regime choose $s^*(t) = R$ (this is true for any $K > 0$) and types sufficiently far from it have a unique inner solution. For the subrange where all follow the regime we have $ds^*/dt = 0$. For the subrange with inner solutions using $\beta = 1$ and $\alpha > 1$ in equation (9) implies that $ds^*/dt = 1$ and hence $|s^*(t) - R|$ is increasing in the distance to the regime.

$1 = \alpha < \beta$: Solving for the range $t > R$ and then using symmetry around R, it is easy to verify that types sufficiently close to the regime choose $s^*(t) = t$, while types sufficiently far from the regime choose the same inner solution s, s.t. $P'(|s - R|) = 1$ $(= D')$. It thus follows that $|s^*(t) - R|$ is first increasing in the distance from the regime and then it stays constant.

A.3 Proof of Proposition 1

A.3.1 Part 1

We start by showing that initially – i.e., in the steady state – the most dissenting types are extremists (i.e., $\max |s^*(t) - R|$ is achieved for $t = \arg\max_t |t - R|$). For $\alpha \leq 1$ this follows immediately from Section A.2.1. If instead $\alpha > 1$, we know from Section A.2.2 that if the range of types is not sufficiently broad, then $s^*(t) = R$ for everyone hence the claim trivially holds. Otherwise, if the range of types is sufficiently broad so that types sufficiently far from R have an inner solution, Section A.2.2 further tells us that $s^*(t)$ is increasing in the subrange of types with inner solutions, implying that $\max |s^*(t) - R|$ is achieved for $t = \arg\max_t |t - R|$.

To see that, as the revolution evolves, more moderate types join, note first that during the revolution K decreases. Sections A.2.1 and A.2.2 tell us that, when $\beta < \alpha$, types sufficiently close to the regime (moderates) support the regime. Consider now the cutoff type at time i, who supports the regime $(s^*(t) = R)$ but is indifferent between R and some $s \neq R$ ($s = t$ in the case of $\alpha \leq 1$ and some inner solution in the case of $\alpha > 1$). This means that, for this type, the difference between the two alternative solutions in terms of regime sanctioning P exactly cancels out with the difference between the two alternative solutions in terms of the discomfort D. At time $i + 1$ the regime becomes weaker, hence the difference between the two alternative solutions in terms of regime sanctioning P must become smaller than the difference between the two alternative solutions in terms of the discomfort D, implying that this type will stop supporting the regime and instead join the revolution.

A.3.2 Part 2

We start by showing that initially – i.e., in the steady state – the most dissenting types are moderates (i.e., $\max |s^*(t) - R|$ is achieved for $t \neq \arg\max_t |t - R|$). If $\beta \leq 1$, we know from Section A.2.1 that there exists a distance from the regime, $\Delta = K^{\frac{1}{\alpha - \beta}}$, such that a type at that distance chooses $s^*(t) = t$ and hence has $|t - R| = \Delta$, while any type further away from R has $|t - R| = 0$. Given that, in a steady state with a regime, Δ must be smaller than $\max_t |t - R|$ (as otherwise $s^*(t) = t$ for everyone hence the regime does not exist), this immediately implies that $\max |s^*(t) - R| = \Delta$ is achieved for $t = R \pm \Delta \neq \arg\max_t |t - R|$. Alternatively, if $\beta > 1$, we know from Section A.2.3 that if the range of types is not sufficiently broad, then $s^*(t) = t$ for everyone hence a regime does not exist. If a regime exists it therefore must be that types sufficiently far from R have an inner solution. Moreover, Section A.2.3 further tells us that $s^*(t)$ is decreasing in the subrange of types with inner solutions, implying that $\max |s^*(t) - R|$ is achieved for $t \neq \arg\max_t |t - R|$.

To see that, as the revolution evolves, more extreme types (compared to $\arg\max_t |s^*(t) - R|$ at the steady state) dissent the most, note first that during the revolution K decreases. This implies that the most dissenting type at time $i+1$ (who, at this point in time, chooses $s^*(t) = t$) must have had a different solution at time i ($s_i^*(t) = R$ if $\beta \leq 1$, or an inner solution if $\beta > 1$), implying that she is further away from the regime (= a more extreme type) than the type who was most dissenting at time i (who herself is more extreme than the one most dissenting at time $i-1$ and so on until we reach the steady state).

A.3.3 Part 3

That initially – i.e., in the steady state – the most dissenting types are extremists (i.e., $\max |s^*(t) - R|$ is achieved for $t = \arg\max_t |t - R|$), follows immediately from Section A.2.4, where we show that $|s^*(t) - R|$ is increasing in the distance to the regime. During the revolution K decreases, making any type with an inner solution choose a new stance further away from the regime. In the special case where $1 = \beta < \alpha$ and we start with a steady state were all follow the regime, the revolution will be triggered by someone stopping to follow it, where the analysis in Section A.2.4 implies that these will be the types furthest away from the regime, and they will have inner solutions, hence, again, will gradually choose solutions further and further away from the regime.

A.4 Extremists starting with extreme dissent: $\beta < \alpha \leq 1$

First note from Sections A.2.1 that $s^*(t)$ is uniquely defined for all types (except for at most one, infinitesimal type who may be indifferent between the two corners). Hence for any K and hence A there exists a unique set of stances. This means that, in the upcoming analyses of the steady states it is sufficient to look at situations where $A_{i+1} = f(A_{i+1})$.

A.4.1 The phase diagram

We start by analyzing the behavior of A_{i+1} as a function of A_i, as depicted graphically in the phase diagram (Figure 3). As will be proved below, the phase diagram contains at most four parts, corresponding to the following cases (described from left to right in the diagram):

1. A sufficiently small A_i, which produces $A_{i+1} = 0$, indicating the case where $s_{i+1}(t) = t$ $\forall t$, and the phase diagram is flat.

2. A bit larger A_i, for which types far from the regime on both sides of it choose $s_{i+1}(t) = t$, while for the rest $s_{i+1}(t) = R$.

3. An even larger A_i, for which only types far from the regime on the far side of it choose $s_{i+1}(t) = t$, while for the rest $s_{i+1}(t) = R$.

4. A sufficiently large A_i, which produces $A_{i+1} = 1$, reflecting the case where $s_{i+1}(t) = R$ $\forall t$, and the phase diagram is flat.

We now prove that this is indeed the shape of the phase diagram. The analytical properties of $A_{i+1} = f(A_i)$ and of the individuals' behavior are summarized in the following lemma.

Lemma 1 *Suppose $\beta < \alpha \leq 1$. Then:*

1. *$A_{i+1} = f(A_i)$ is continuous and increasing in A_i.*
2. *There exists an $\varepsilon \geq 0$ such that $A_{i+1} = f(A_i) = 0$ for all $A_i \leq \varepsilon$. $\varepsilon = 0$ iff $|R| = 0$.*
3. *When $R = 0$ then $f(A_i)$ is convex for $A_i > 0$.*
4. *When $R \neq 0$ then for $A_i > \varepsilon$, $f(A_i)$ is convex initially. If $R \in [-1, -1/2[$, it stays convex throughout. Otherwise, if $R \in [-1/2, 0]$, then at the A_i corresponding to $\Delta = 1 + R$ the slope of $f(A_i)$ discontinuously decreases and $f(A_i)$ is convex thereafter until either $f(A_i)$ or A_i reaches 1.*
5. *Holding all else fixed, $f(A_i)$ is weakly decreasing in $|R|$.*
6. *Holding all else fixed, $f(A_i)$ is weakly increasing in $\bar{K}$.*
7. *The unstable steady states (A_{uss}) are increasing in $|R|$ while the stable steady states (A_{ss}) are (weakly) decreasing in $|R|$.*
8. *There exists a $\bar{K}_{c1}$ such that a stable steady state with a regime and $A_{ss} > 0$ exists iff $\bar{K} > \bar{K}_{c1}$.*
9. *$\bar{K}_{c1}$ is increasing in $|R|$.*

Proof. *From Section A.2.1 we know that (for sufficiently large K) there is a cutoff distance Δ between regime conformers (within the cutoff) and those speaking their minds (beyond the cutoff) s.t. $\Delta \equiv K^{\frac{1}{\alpha-\beta}} = \left(\bar{K}A\right)^{\frac{1}{\alpha-\beta}}$. Suppose, without loss of generality, that $R \leq 0$. If $\Delta \leq 1 - |R|$ (which is the distance from the regime to the closest edge of the type distribution), we have by equation (8)*

$$\begin{aligned}\Psi\left(s_i^*; R, A_i\right) &= \int_{-1}^{R-\Delta_i} (R - t)\, dt + \int_{R+\Delta_i}^{1} (t - R)\, dt \\ &= \ldots = R^2 - \Delta_i^2 + 1\end{aligned}$$

while if $\Delta > 1 - |R|$ we have

$$\begin{aligned}\Psi\left(s_i^*; R, A_i\right) &= \int_{R+\Delta_i}^{1} (t - R)\, dt = \ldots \\ &= \frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2.\end{aligned}$$

Hence we get

$$\Psi\left(s_i^*; R, A_i\right) = \begin{cases} R^2 - \Delta_i^2 + 1 \text{ when } 0 \leq \Delta_i \leq 1 + R \\ \frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2 \text{ when } 1 + R < \Delta_i < 1 - R \\ 0 \text{ when } \Delta_i \geq 1 - R \end{cases}.$$

Noting that $A_{i+1} = 0$ by construction whenever $\Psi\left(s_i^*; R, A_i\right) \geq 1$, we start by checking whether this inequality may hold in the first region of $\Psi\left(s_i^*; R, A_i\right)$.

$$\begin{aligned} 1 &\leq R^2 - \Delta_i^2 + 1 \\ &\Leftrightarrow \Delta_i \leq -R. \end{aligned}$$

If $R \in [-1, -1/2]$, this inequality holds throughout the first region (i.e. for any $0 \leq \Delta_i \leq 1 + R$), which means that $\Psi\left(s_i^*; R, A_i\right) \geq 1$ may hold also for some Δ_i in the middle region. Checking when this happens we get

$$\begin{aligned} \frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2 &= 1 \Rightarrow ... \Rightarrow \\ \Delta_i &= \sqrt{\left(R - \left(1 + \sqrt{2}\right)\right)\left(R - \left(1 - \sqrt{2}\right)\right)}, \end{aligned}$$

which does fall within the range $1 + R < \Delta_i < 1 - R$ for $R \in [-1, -1/2]$. Thus, in this case where $R \in [-1, -1/2[$ we get

$$A_{i+1} \equiv f\left(R, A_i\right) = \begin{cases} 0 \text{ when } \Delta_i \leq -R \\ 1 - \left(\frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2\right) \text{ when } -R < \Delta_i < 1 - R \\ 1 \text{ when } 1 - R < \Delta_i \end{cases}$$

Otherwise, for $R \in [-1/2, 0]$, $\Psi\left(s_i^*; R, A_i\right) \geq 1$ may hold only in the first region, and we get

$$A_{i+1} \equiv f\left(R, A_i\right) = \begin{cases} 0 \text{ when } 0 \leq \Delta_i \leq -R \\ 1 - \left(R^2 - \Delta_i^2 + 1\right) \text{ when } -R < \Delta_i \leq 1 + R \\ 1 - \left(\frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2\right) \text{ when } 1 + R < \Delta_i < 1 - R \\ 1 \text{ when } 1 - R \leq \Delta_i \end{cases}. \qquad (14)$$

These four regions correspond to the four schematically described above. As the three-regions phase diagram for $R \in [-1, -1/2[$ can be seen as a degenerate version of the four-regions phase diagram for $R \in [-1/2, 0]$, we will continue the analysis only for the latter case. Recalling that

$$\Delta_i = \left(\bar{K} A_i\right)^{\frac{1}{\alpha - \beta}}, \qquad (15)$$

and noting that this expression is monotonically increasing in A_i for $\beta < \alpha$, we get that $A_{i+1} = 0$ for any $A_i \leq \varepsilon \equiv \frac{(-R)^{\alpha - \beta}}{\bar{K}}$, where $\varepsilon \geq 0$ and $\varepsilon = 0$ iff $|R| = 0$. As Figure 3 shows and will now be proved, the two middle regions are convex. Using (14) and (15)

$$\begin{aligned} \frac{df}{dA_i} &= \begin{cases} \frac{2}{\alpha - \beta}\Delta_i^2 A_i^{-1} \text{ when } \Delta_i \leq 1 + R \\ \frac{1}{\alpha - \beta}\Delta_i^2 A_i^{-1} \text{ when } 1 + R < \Delta_i < 1 - R \end{cases} > 0 \\ \frac{d^2 f}{dA_i^2} &= \begin{cases} \frac{2}{\alpha - \beta}\frac{\Delta_i^2}{A_i^2}\left(\frac{2}{\alpha - \beta} - 1\right) \text{ when } \Delta_i \leq 1 + R \\ \frac{1}{\alpha - \beta}\frac{\Delta_i^2}{A_i^2}\left(\frac{2}{\alpha - \beta} - 1\right) \text{ when } 1 + R < \Delta_i < 1 - R \end{cases} > 0 \end{aligned}$$

since $\alpha - \beta \in (0, 1)$. Thus, for $R \in [-1/2, 0]$ the function f has a kink at $\Delta_i = 1 + R$ with a lower slope after the kink. These properties imply that the phase-diagram is flat at zero, convexly increasing, then has a downward kink and is convexly increasing after. This proves parts (1)-(4). There are at most two stable steady states, one at $A_i = 1$ and one interior.

Since $A_{i+1} = f(A_i)$ is flat at zero it means that the first intersection is unstable, the next is stable, next unstable and next stable. Using (14) and (15)

$$\frac{df}{dR} = \left\{ \begin{array}{c} -2R \textit{ when } -R < \Delta_i \leq 1+R \\ 1-R \textit{ when } 1+R < \Delta_i < 1-R \\ 0 \textit{ otherwise} \end{array} \right. \geq 0$$

since $R \leq 0$, proving part (5). Furthermore,

$$\frac{df}{d\bar{K}} = \left\{ \begin{array}{c} \frac{2}{\alpha-\beta}\Delta_i^2/\bar{K} \textit{ when } -R < \Delta_i \leq 1+R \\ \frac{1}{\alpha-\beta}\Delta_i^2/\bar{K} \textit{ when } 1+R < \Delta_i < 1-R \\ 0 \textit{ otherwise} \end{array} \right. \geq 0,$$

proving part (6). These results imply that the unstable steady states (A_{uss}) are increasing in $|R|$ and decreasing in $\bar{K}$. The stable steady states (A_{ss}) are (weakly) decreasing in $|R|$ and (weakly) increasing in $\bar{K}$. This proves part (7).

Since it was shown that the phase diagram $A_{i+1} = f(A_i)$ starts below the 45-degree line, if follows that a stable steady state exists if A_{i+1} crosses the 45-degree line at least once. For this to happen, one of the following conditions should hold:

1. *The kink is above the 45 degree line: $A_{i+1}|_{\Delta_i=1+R} \geq A_i|_{\Delta_i=1+R} \Leftrightarrow$ {Using (14) and (15)} $\Leftrightarrow 1+2R \geq (1+R)^{\alpha-\beta}/\bar{K}$. As the RHS is positive, this inequality can hold only if*

   $$R \in [-1/2, 0] \textit{ and } \bar{K} \geq \frac{(1+R)^{\alpha-\beta}}{1+2R}$$

2. *$A_{i+1}(1) = 1$ (i.e., $1-R \leq \Delta_i$ when $A_i = 1$)$\Leftrightarrow$ {Using (15)} $\Leftrightarrow 1-R \leq \bar{K}^{\frac{1}{\alpha-\beta}} \Leftrightarrow$*

   $$\bar{K} \geq (1-R)^{\alpha-\beta}$$

   Denote the smallest $\bar{K}$ fulfilling one of these conditions by $\bar{K}_{c1}$. Thus follows part (8), and it can be verified that $\bar{K}_{c1}$ is increasing in $|R|$ (proving part (9)).

■

A.4.2 Proof of Proposition 2

Part (1): follows from parts (8) and (9) of Lemma 1.

Part (2): We first remind that a revolution is defined as a dynamic process where the approval is converging to a new, lower steady state. From this definition it directly follows that a negative shock to the approval of a regime in a stable steady state (with approval A_{ss}), such that the size of the shock is larger than $|A_{ss} - A_{uss}|$ (where A_{uss} is the approval in the closest unstable steady state to the left), would result in a revolution. A negative shock to the force ($\bar{K}$) of the regime reduces A_{i+1} (part (6) of Lemma 1), and in particular if the shock is such that $\bar{K}$ goes below $\bar{K}_{c1}$, A converges to zero and the regime completely falls (part (8) of Lemma 1). Finally, implementation of unpopular policies means that $|R|$ increases, and as a result the approval of the regime decreases (part (5) of Lemma 1), and in particular a revolution would start if the approval decreases sufficiently to eliminate the pre-existing stable steady state.

Part (3): (a) follows directly from part (1) of Proposition 1. (b) follows from the facts that (i) before the revolution everyone fully supports the regime at least on one side of it (as A_{ss} can only be in the third or fourth region of equation (14) – see Figure 3) and (ii) Δ_i starts above $1 + R$ (where $s(t)$ might be different than R only on one side of the regime).

Part (4): follows from part (1) of Proposition 1 and from the fact that dissenters speak their minds ($s(t) = t$).

A.5 Moderates starting with moderate dissent: $\alpha < \beta \leq 1$

First note from Sections A.2.1-A.2.4 that $s^*(t)$ is uniquely defined for all types (except for at most one, infinitesimal type who may be indifferent between the two corners). Hence for any K and hence A there exists a unique set of stances. This means that, in the upcoming analyses of the steady states it is sufficient to look at situations where $A_{i+1} = f(A_{i+1})$.

A.5.1 The phase diagram

We start by analyzing the behavior of A_{i+1} as a function of A_i, as depicted graphically in the phase diagram (Figure 6). As will be proven below, the phase diagram contains at most three parts, corresponding to the following cases (described from left to right in the diagram):

1. A sufficiently small A_i, which produces $A_{i+1} = 0$, indicating the case where $s_{i+1}(t) = t$ $\forall t$, and the phase diagram is flat.

2. A smaller A_i, for which types far from the regime on one side of it choose $s_{i+1}(t) = R$, while for the rest $s_{i+1}(t) = t$.

3. A sufficiently large A_i, for which types far from the regime on both sides of it choose $s_{i+1}(t) = R$, while for the rest $s_{i+1}(t) = t$.

We now prove that this is indeed the shape of the phase diagram. The analytical properties of $A_{i+1} = f(A_i)$ and of the individuals' behavior are summarized in the following lemma.

Lemma 2 *Suppose $\alpha < \beta \leq 1$. Then:*

1. *$A_{i+1} = f(A_i)$ is continuous and increasing in A_i.*

2. *There exists an $\varepsilon > 0$ such that $A_{i+1} = f(A_i) = 0$ for all $A_i \leq \varepsilon$.*

3. *When $R = 0$ then $f(A_i)$ is concave for $A_i > \varepsilon$.*

4. *When $R \neq 0$ then for $A_i > \varepsilon$, $f(A_i)$ is concave initially. At the A_i implied by $\Delta_i = 1 - |R|$ the slope of $f(A_i)$ discontinuously increases and $f(A_i)$ is concave thereafter until A_i reaches 1.*

5. *Holding all else fixed, $f(A_i)$ is weakly increasing in $|R|$.*

6. *Holding all else fixed, $f(A_i)$ is weakly increasing in $\bar{K}$.*

7. *The unstable steady states (A_{uss}) are weakly decreasing in $|R|$ while the stable steady states (A_{ss}) are weakly increasing in $|R|$.*

8. *$f(1) < 1$.*

9. *There exists a $\bar{K}_{c2}$ such that a stable steady state with a regime and $A_{ss} > 0$ exists iff $\bar{K} > \bar{K}_{c2}$.*

10. *$\bar{K}_{c2}$ is weakly decreasing in $|R|$.*

Proof. *From Section A.2.1 we know that (for sufficiently large K) there is a cutoff distance Δ between regime conformers ($|t - R| > \Delta$) and those speaking their minds ($|t - R| \leq \Delta$) such that $\Delta \equiv K^{\frac{1}{\alpha-\beta}} = \left(\bar{K}A\right)^{\frac{1}{\alpha-\beta}}$. Suppose, without loss of generality, that $R \leq 0$. If $\Delta \leq 1-|R|$ (which is the distance from the regime to the closest edge of the type distribution), we have by equation (8)*

$$\begin{aligned}\Psi\left(s_i^*; R, A_i\right) &= \int_{R-\Delta_i}^{R} (R-\tau)\, d\tau + \int_{R}^{R+\Delta_i} (\tau - R)\, d\tau \\ &= \Delta_i^2\end{aligned}$$

while if $\Delta > 1 - |R|$ we have

$$\begin{aligned}\Psi\left(s_i^*; R, A_i\right) &= \int_{-1}^{R} (R-\tau)\, d\tau + \int_{R}^{R+\Delta_i} (\tau - R)\, d\tau \\ &= \frac{1}{2}(1+R)^2 + \frac{1}{2}\Delta_i^2\end{aligned}$$

Hence we get

$$\Psi\left(s_i^*; R, A_i\right) = \begin{cases} \Delta_i^2 \text{ when } 0 \leq \Delta_i \leq 1+R \\ \frac{1}{2}(1+R)^2 + \frac{1}{2}\Delta_i^2 \text{ when } 1+R < \Delta_i < 1-R \\ 1 + R^2 \text{ when } 1 - R \leq \Delta_i \end{cases}$$

noting that $\Psi\left(s_i^*; R, A_i\right)$ might equal 1 only in the middle range (unless $R = 0$), and in particular when

$$\begin{aligned} 1 &= \frac{1}{2}(1+R)^2 + \frac{1}{2}\Delta_i^2 \\ &\Leftrightarrow\; 1 - R^2 - 2R = \Delta_i^2 \end{aligned}$$

we get by (7) that

$$A_{i+1} \equiv f\left(R, A_i\right) = \begin{cases} 1 - \Delta_i^2 \text{ when } 0 \leq \Delta_i \leq 1+R \\ 1 - \left(\frac{1}{2}(1+R)^2 + \frac{1}{2}\Delta_i^2\right) \text{ when } 1+R < \Delta_i < \sqrt{1-R^2-2R} \\ 0 \text{ when } \sqrt{1-R^2-2R} \leq \Delta_i \end{cases} . \tag{16}$$

These three regions correspond to the three schematically described above. Recalling that

$$\Delta_i = \left(\bar{K}A_i\right)^{\frac{1}{\alpha-\beta}}, \tag{17}$$

and noting that this expression is monotonically deceasing in A_i for $\alpha < \beta$, we get that $A_{i+1} = 0$ for any $A_i \leq \varepsilon \equiv \frac{\left(\sqrt{1-R^2-2R}\right)^{\alpha-\beta}}{\bar{K}}$, where $\varepsilon > 0$. As Figure 6 shows and will now *be proved, the two regions in which* $A_{i+1} \neq 0$ *are concave. Using (16) and (17) we get*

$$\begin{aligned} \frac{df}{dA_i} &= \left\{ \begin{array}{c} -\frac{2}{\alpha-\beta}\Delta_i^2 A_i^{-1} \textit{ when } \Delta_i \leq 1+R \\ -\frac{1}{\alpha-\beta}\Delta_i^2 A_i^{-1} \textit{ when } 1+R < \Delta_i < \sqrt{1-R^2-2R} \end{array} \right. > 0 \\ \frac{d^2 f}{dA_i^2} &= \left\{ \begin{array}{c} -\frac{2}{\alpha-\beta}\frac{\Delta_i^2}{A_i^2}\left(\frac{2}{\alpha-\beta}-1\right) \textit{ when } \Delta_i \leq 1+R \\ -\frac{1}{\alpha-\beta}\frac{\Delta_i^2}{A_i^2}\left(\frac{2}{\alpha-\beta}-1\right) \textit{ when } 1+R < \Delta_i < \sqrt{1-R^2-2R} \end{array} \right. < 0 \end{aligned}$$

since $\alpha-\beta \in (-1,0)$. *Thus, the function* f *has a kink at* $\Delta_i = 1+R$ *with a bigger slope after the kink (note that small values of* Δ *correspond to high approval and large values correspond to low approval). These properties imply that the phase-diagram is first flat, then concavely increasing, then has an upward kink and is concavely increasing thereafter. This proves parts (1)-(4). There are at most two (interior) stable steady states. Since* $A_{i+1} = f(A_i)$ *is flat at zero, it means that the first intersection is unstable, the next is stable, next unstable and next stable.*

$$\frac{df}{dR} = \left\{ \begin{array}{c} -1-R \textit{ when } 1+R < \Delta_i < \sqrt{1-R^2-2R} \\ 0 \textit{ otherwise} \end{array} \right. \leq 0$$

since $R \geq -1$, *proving part (5). Furthermore,*

$$\frac{df}{d\bar{K}} = \left\{ \begin{array}{c} -\frac{2}{\alpha-\beta}\Delta_i^2/\bar{K} \textit{ when } -R < \Delta_i \leq 1+R \\ -\frac{1}{\alpha-\beta}\Delta_i^2/\bar{K} \textit{ when } 1+R < \Delta_i < \sqrt{1-R^2-2R} \\ 0 \textit{ otherwise} \end{array} \right. \geq 0,$$

proving part (6). These results imply that the unstable steady states (A_{uss}) *are decreasing in* $|R|$ *and in* $\bar{K}$. *The stable steady states* (A_{ss}) *are increasing in* $|R|$ *and in* $\bar{K}$. *This proves part (7). When* $A_i = 1$ *we get by (17) that* Δ_i *is strictly positive, hence, by (16),* $A_{i+1} < 1$, *which proves part (8). This further implies, together with the fact that the phase diagram* $A_{i+1} = f(A_i)$ *starts below the 45 degree line, that a necessary and sufficient condition for the existence of a stable steady state is that* f *crosses (and not just touches) the 45-degree line. Now, note that for*

$$\bar{K} = \frac{(1+R)^{\alpha-\beta}}{1-(1+R)^2} \tag{18}$$

we get that the kink is exactly on the 45-degree line, because this yield

$$1-(1+R)^2 = (1+R)^{\alpha-\beta}/\bar{K} \Rightarrow \{\textit{using (16) and (17)}\} \Rightarrow A_{i+1}|_{\Delta_i=1+R} = A_i|_{\Delta_i=1+R},$$

in which case a stable steady state exists. Next, part (6) implies that f *is weakly increasing in* K, *so that if for a certain* K^* *a stable steady state exists, then a stable steady state exists for any* $K > K^*$. *Denote the smallest* $\bar{K}$ *for which* f *touches the 45-degree line (given by (18)) by* $\bar{K}_{c2}$. *Thus follows part (9), and part (10) follows from the fact that* f *increases in* $\bar{K}$ *and* $|R|$ *(by parts (5) and (6)).* ■

A.5.2 Proof of Proposition 3

Part (1) follows from Lemma 2 parts (9) and (10).

Part (2): We first remind that a revolution is defined as a dynamic process where the approval is converging to a new, lower steady state. From this definition it directly follows that a negative shock to the approval of a regime in a stable steady state (with approval A_{ss}), such that the size of the shock is larger than $|A_{ss} - A_{uss}|$ (where A_{uss} is the approval in the closest unstable steady state to the left), would result in a revolution. A negative shock to the force of the regime reduces A_{i+1} (part (6) of Lemma 2), and in particular if the shock is such that $\bar{K}$ goes below $\bar{K}_{c2}$, then A_i converges to zero and the regime completely falls (part (9) of Lemma 2). Finally, implementation of popular policies means that $|R|$ decreases, and as a result the approval of the regime decreases as well (part (5) of Lemma 2), and in particular a revolution would start if the approval decreases sufficiently to eliminate the pre-existing stable steady state.

Part (3): (a) follows directly from part (2) of Proposition 1. (b) follows from the fact that dissent at time i comes from people within the cutoff Δ_i, and for any R s.t. $|R| \neq 1$ this implies dissent on both sides of the regime.

Part (4): follows from the facts that (i) dissent at time i comes from people within the cutoff Δ_i (see part (2) of Proposition 1), (ii) Δ_i increases as A_i decreases during the revolution, and (iii) dissenters speak their minds ($s(t) = t$).

A.6 Extremists starting with moderate dissent: $\alpha > 1$, $\beta \geq 1$

When $\alpha > 1$, $\beta \geq 1$, every type $t > R$ has a unique inner solution $s^*(t) \in]R, t[$ and every type $t < R$ has a unique inner solution $s^*(t) \in]t, R[$, with this solution being determined by equation (12) (see Section A.2.4). This means that for any K and hence A there exists a unique set of stances implying that, in the upcoming analyses of the steady states it is sufficient to look at situations where $A_{i+1} = f(A_{i+1})$.

Substituting variables to $\sigma \equiv |s^*(t) - R|$ and $\tau \equiv |t - R|$ yields

$$\begin{aligned} K_i \beta \sigma^{\beta-1} &= \alpha (\tau - \sigma)^{\alpha-1} \\ &\Leftrightarrow \tau = \sigma + \left(\frac{K_i \beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \sigma^{\frac{\beta-1}{\alpha-1}}. \end{aligned} \tag{19}$$

We turn now to calculating $\Psi(s_i^*; R, A_i)$. To do that, we first remind that $\Psi(s_i^*; R, A_i)$ is the sum of deviations from R (i.e. the sum of $\sigma(t)$ $\forall t$). Hence, it equals the area under the graph of $\sigma(t)$. Now, since σ is an implicit function of t (and of τ), it is difficult to compute the integral of $\sigma(\tau)$ (= the area under $\sigma(t)$). Instead, it is easier to compute it using the explicit expression of $\tau(\sigma)$ in (19). Noting that, at each side of R, σ is monotonous in t, we can substitute the calculation of the area under $\sigma(\tau)$ for positive τ with a calculation of the area above $\tau(\sigma)$ and below a horizontal line at the value $1 - R$ (which is max τ), and the calculation of the area under $\sigma(\tau)$ for negative τ with a calculation of the area below $\tau(\sigma)$ and above a horizontal line at the value $-(1 + R)$ (which is min τ).[27] Finally, using the

[27] To see this it is easiest to draw a generic increasing function $\sigma(\tau)$ between 0 and $1 + R$ and note, by turning the drawing 90 degrees, that the area it creates is the same as the area given by $1 + R - \tau(\sigma)$ with boundaries $\sigma(0)$ and $\sigma(1 + R)$.

symmetry of $\sigma(\tau)$ around 0 we can substitute $\int_{-(1+R)}^{0}\sigma(\tau)\,d\tau$ with $\int_{0}^{1+R}\sigma(\tau)\,d\tau$ to get

$$\begin{aligned}
\Psi\left(s_i^*;R,A_i\right) &= \int_0^{1+R}\sigma(\tau)\,d\tau+\int_0^{1-R}\sigma(\tau)\,d\tau \\
&= \int_0^{\check{\sigma}\equiv\sigma(1+R)}\left[(1+R)-\tau(\sigma)\right]d\sigma+\int_0^{\hat{\sigma}\equiv\sigma(1-R)}\left[(1-R)-\tau(\sigma)\right]d\sigma \\
&= \int_0^{\check{\sigma}\equiv\sigma(1+R)}\left[(1+R)-\sigma-\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}\right]d\sigma \\
&\quad+\int_0^{\hat{\sigma}\equiv\sigma(1-R)}\left[(1-R)-\sigma-\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}\right]d\sigma \\
&= (1+R)\,\check{\sigma}-\frac{\check{\sigma}^2}{2}+(1-R)\,\hat{\sigma}-\frac{\hat{\sigma}^2}{2}-\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1} \qquad (20)
\end{aligned}$$

The analytical properties of $A_{i+1}=f(A_i)$ and of the individuals' behavior are summarized in the following lemma.

Lemma 3 *Suppose $\alpha>1$, $\beta\geq 1$. Then:*

1. *$A_{i+1}=f(A_i)$ is continuous and increasing in A_i.*
2. *There exists an $\varepsilon\geq 0$ such that $A_{i+1}=f(A_i)=0$ for all $A_i\leq\varepsilon$. $\varepsilon=0$ iff $|R|=0$.*
3. *For $A_i>\varepsilon$, $f(A_i)$ is first convex then concave, or convex throughout, or concave throughout.*
4. *Holding all else fixed, $f(A_i)$ is decreasing in $|R|$.*
5. *Holding all else fixed, $f(A_i)$ is increasing in $\bar{K}$.*
6. *$f(1)<1$.*
7. *There exists a $\bar{K}_{c3}$ such that a stable steady state with a regime and $A_{ss}>0$ exists iff $\bar{K}>\bar{K}_{c3}$.*
8. *$\bar{K}_{c3}$ is increasing in $|R|$.*
9. *There are at most two steady states with $A>0$, where the first is unstable and the second is stable.*
10. *The unstable steady states (A_{uss}) are increasing in $|R|$ while the stable steady states (A_{ss}) are (weakly) decreasing in $|R|$.*

Proof. *To see that part (1) holds, recall that by construction (6) $A=\max\left\{0,1-\Psi\left(s_i^*;R,A_i\right)\right\}$*

and note that

$$\begin{aligned}\frac{d\Psi\left(\sigma_i;R,A_i\right)}{dA_i} &= \left(1+R-\check{\sigma}\right)\frac{d\check{\sigma}}{dA_i}+\left(1-R-\hat{\sigma}\right)\frac{d\hat{\sigma}}{dA_i}-\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-1}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}\\ &\quad-\left(\frac{\bar{K}A_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\left(\check{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\check{\sigma}}{dA_i}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\hat{\sigma}}{dA_i}\right)\\ &= \left(1+R-\check{\sigma}-\left(\frac{\bar{K}A_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\check{\sigma}^{\frac{\beta-1}{\alpha-1}}\right)\frac{d\check{\sigma}}{dA_i}+\left(1-R-\hat{\sigma}-\left(\frac{\bar{K}A_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\hat{\sigma}^{\frac{\beta-1}{\alpha-1}}\right)\frac{d\hat{\sigma}}{dA_i}\\ &\quad-\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-1}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}.\end{aligned}$$

Using $\check{\sigma}$ and $\hat{\sigma}$ in the FOC in (10) we get

$$\alpha\left(1+R-\check{\sigma}\right)^{\alpha-1} = \bar{K}A_i\beta\check{\sigma}^{\beta-1} \tag{21}$$

$$\alpha\left(1-R-\hat{\sigma}\right)^{\alpha-1} = \bar{K}A_i\beta\hat{\sigma}^{\beta-1}. \tag{22}$$

Using these in the previous expression for $\frac{d\Psi(\sigma_i;R,A_i)}{dA_i}$ we get that

$$\frac{d\Psi\left(\sigma_i;R,A_i\right)}{dA_i}=-\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-1}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}<0,$$

hence A_{i+1} is increasing in A_i (continuity follows trivially from the definition of A_{i+1} in (6) and the expression of $\Psi\left(\sigma_i;R,A_i\right)$). When $A_i\to 0$ also $K_i\to 0$ hence $\sigma\left(\tau\right)\to\tau$ for all types. For $K_i=0$ we have $\sigma\left(\tau\right)=\tau$ and $\Psi\left(s_i^*;R,A_i\right)=\frac{(1-R)^2+(1+R)^2}{2}\geq 1$, with equality only for $R=0$. From (6) and (8) it thus follows that $\exists\varepsilon\geq 0$ such that $A_{i+1}=f\left(A_i\right)=0$ for any $A_i\leq\varepsilon$, where $\varepsilon=0$ iff $|R|=0$. This proves part (2). To prove part (3) we differentiate $\Psi\left(\sigma_i;R,A_i\right)$ one more time:

$$\begin{aligned}\frac{d^2\Psi\left(\sigma_i;R,A_i\right)}{dA_i^2} = &-\frac{1}{\alpha-1}\left(\frac{1}{\alpha-1}-1\right)A_i^{\frac{1}{\alpha-1}-2}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}\\ &-\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-1}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\left(\check{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\check{\sigma}}{dA_i}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\hat{\sigma}}{dA_i}\right).\end{aligned}\tag{23}$$

Note that $\frac{d\check{\sigma}}{dA_i}$ and $\frac{d\hat{\sigma}}{dA_i}$ are both negative.[28] *This implies that $\frac{d^2\Psi(\sigma_i;R,A_i)}{dA_i^2}>0$ when $\alpha\geq 2$, hence A_{i+1} is concave.*

[28] This is true since K increases in A_i which in turn makes everyone, including types 1 and -1, choose a solution closer to R.

We now investigate the case $1 < \alpha < 2$. Revisiting equation (19) we can write

$$\begin{aligned} H &= \sigma + \left(\frac{K_i \beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \sigma^{\frac{\beta-1}{\alpha-1}} - \tau = 0 \\ \Rightarrow \quad \frac{d\sigma}{dA_i} &= -\frac{\frac{dH}{dA_i}}{\frac{dH}{d\sigma}} = -\frac{\frac{1}{\alpha-1} A_i^{\frac{1}{\alpha-1}-1} \left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \sigma^{\frac{\beta-1}{\alpha-1}}}{1 + \frac{\beta-1}{\alpha-1} A_i^{\frac{1}{\alpha-1}} \left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \sigma^{\frac{\beta-1}{\alpha-1}-1}} \\ &= \{\textit{using (19)}\} = -\frac{\frac{1}{\alpha-1} A_i^{-1} (\tau - \sigma)}{1 + \frac{\beta-1}{\alpha-1} (\tau - \sigma) \sigma^{-1}}. \qquad (24) \end{aligned}$$

Rewriting (23)

$$\frac{d^2 \Psi(\sigma_i; R, A_i)}{dA_i^2} = -\frac{1}{\alpha-1} A_i^{\frac{1}{\alpha-1}-2} \left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \left[\left(\frac{1}{\alpha-1} - 1\right) \frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1} + \hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1} + A_i \left(\check{\sigma}^{\frac{\beta-1}{\alpha-1}} \frac{d\check{\sigma}}{dA_i} + \hat{\sigma}^{\frac{\beta-1}{\alpha-1}} \frac{d\hat{\sigma}}{dA_i}\right)\right]$$

Using the FOC $\left(\frac{K_i \beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \sigma^{\frac{\beta-1}{\alpha-1}} = \tau - \sigma$ and (24) we get

$$\begin{aligned} \frac{d^2 \Psi(\sigma_i; R, A_i)}{dA_i^2} &= -\frac{1}{\alpha-1} A_i^{-2} \left[\sigma(\tau-\sigma)\left(\frac{1}{\alpha-1} - 1\right) \frac{1}{\frac{\beta-1}{\alpha-1}+1} - (\tau-\sigma) \frac{\frac{1}{\alpha-1}(\tau-\sigma)}{1 + \frac{\beta-1}{\alpha-1}\frac{\tau-\sigma}{\sigma}}\right]\Bigg|_{\tau=1+R} \qquad (25) \\ & \quad -\frac{1}{\alpha-1} A_i^{-2} \left[\sigma(\tau-\sigma)\left(\frac{1}{\alpha-1} - 1\right) \frac{1}{\frac{\beta-1}{\alpha-1}+1} - (\tau-\sigma) \frac{\frac{1}{\alpha-1}(\tau-\sigma)}{1 + \frac{\beta-1}{\alpha-1}\frac{\tau-\sigma}{\sigma}}\right]\Bigg|_{\tau=1-R}. \end{aligned}$$

Note that

$$\begin{aligned} & \sigma(\tau-\sigma)\left(\frac{1}{\alpha-1} - 1\right) \frac{1}{\frac{\beta-1}{\alpha-1}+1} - (\tau-\sigma) \frac{\frac{1}{\alpha-1}(\tau-\sigma)}{1 + \frac{\beta-1}{\alpha-1}\frac{\tau-\sigma}{\sigma}} \\ = \; & \sigma(\tau-\sigma)\left[\frac{2-\alpha}{\alpha-1}\frac{\alpha-1}{\beta+\alpha-2} - \frac{\frac{1}{\alpha-1}(\tau-\sigma)}{\sigma + \frac{\beta-1}{\alpha-1}(\tau-\sigma)}\right] \\ = \; & \sigma(\tau-\sigma)\left[\frac{2-\alpha}{\beta+\alpha-2} - \frac{\frac{\tau-\sigma}{\tau}}{(\alpha-1)\frac{\sigma}{\tau} + (\beta-1)\frac{\tau-\sigma}{\tau}}\right], \end{aligned}$$

where $\frac{2-\alpha}{\beta+\alpha-2} > 0$ for $1 \leq \alpha < 2$ and $\frac{\frac{\tau-\sigma}{\tau}}{(\alpha-1)\frac{\sigma}{\tau}+(\beta-1)\frac{\tau-\sigma}{\tau}}$ is positive and increasing in the relative step that type t takes toward the regime, $\frac{\tau-\sigma}{\tau} \in]0,1[$. Moreover, for any τ and any α s.t. $1 \leq \alpha < 2$, the expression in the squared brackets goes from positive to negative as the relative step $\frac{\tau-\sigma}{\tau}$ grows from 0 to 1. It can further be verified that $\frac{\tau-\sigma}{\tau}$ increases in A_i (because an increase in A_i implies that the regime is stronger and so one needs to accommodate more to R). Returning now to (25) and noting that $\frac{d^2\Psi(\sigma_i;R,A_i)}{dA_i^2}$ has the opposite sign of the squared brackets, we get that, as A_i increases, $\frac{d^2\Psi(\sigma_i;R,A_i)}{dA_i^2}$ either keeps its sign or changes sign once, from negative to positive. Finally, since $A_{i+1} = \max\{0, 1 - \Psi_i(\sigma_i; R, A_i)\}$, we get that $A_{i+1}(A_i)$ is first convex then concave, or convex throughout, or concave throughout, which proves part (3).

Differentiating equation (20) w.r.t. R and then using (21) and (22) yields

$$\frac{d\Psi\left(s_i^*;R,A_i\right)}{dR}=\check{\sigma}-\hat{\sigma}\leq 0$$

(by the monotonicity of $\sigma\left(\tau\right)$), implying that A_{i+1} decreases in $|R|$, which proves part (4). Next, differentiating equation (20) by $\bar{K}$ and then using (21) and (22) yields

$$\frac{d\Psi\left(s_i^*;R,A_i\right)}{d\bar{K}}=-\frac{1}{\alpha-1}\bar{K}^{\frac{1}{\alpha-1}-1}\left(\frac{A_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}<0,$$

hence $f\left(A_i\right)$ is increasing in $\bar{K}$, which proves part (5). Part (6) follows from the fact that all types always have inner solutions (for finite $\bar{K}$) to the optimization problem, hence $A_{i+1}=f\left(1\right)$ never reaches 1. This further implies, together with the fact that the phase diagram $A_{i+1}=f\left(A_i\right)$ starts below the 45 degree line, that a necessary and sufficient condition for the existence of a stable steady state is that this diagram crosses (and not just touches) the 45-degree line. Now, fix α,β and R, and set $\bar{K}$ to be sufficiently large such that for $\max\tau=1-R$ and $A_i=1/2$, the value of σ which solves equation (19) is smaller than $1/2$. The strict monotonicity of $\sigma\left(\tau\right)$ implies then that the total sum of deviations from the regime ($\Psi\left(s_i^*;R,A_i\right)$) will be smaller than $1\cdot 1/2$, and so $A_{i+1}>1-1/2=1/2=A_i$. In other words, at $A_i=1/2$ the phase diagram is above the 45-degree line, and together with parts (2) and (6) we get that (for $R\neq 0$) the phase diagram crosses the 45-degree line at least twice, and one of these crossing points must be a stable steady state.[29] Furthermore, this happens for finite $\bar{K}$. Together with this result, part (5) implies that $f\left(A_i\right)$ is increasing in $\bar{K}$, so that if for a certain K^* a stable steady state exists, then a stable steady state exists for any $K>K^*$. Denote the smallest $\bar{K}$ for which the diagram touches the 45-degree line by $\bar{K}_{c3}$. Thus follows part (7), and part (8) follows from the fact that $f\left(A_i\right)$ decreases in $|R|$ and decreases in $\bar{K}$ (by part (4) and (5)). Given that the phase diagram starts and ends below the 45-degree line (except for one special case – see previous footnote), it cannot cross this line if it is convex throughout, which (by part (3)) implies that, for $A_i>\varepsilon$, it must be either concave throughout or first convex and then concave. In both cases this leads to at most two crossing points of the 45-degree line, the first from below (hence unstable) and the second from above (hence stable). This proves part (9). Increasing $|R|$ reduces A_{i+1} (by part (4)), and so the new crossing points, if they still exist, lie in the range that previously was above the 45-degree line, $]A_{uss},A_{ss}[$, which means that A_{uss} increases while A_{ss} decreases. This proves part (10).[30] ■

A.6.1 Proof of Proposition 4

Part (1) follows from Lemma 3 parts (7) and (8).

Part (2): We first remind that a revolution is defined as a dynamic process where the approval is converging to a new, lower steady state. From this definition it directly follows that a negative shock to the approval of a regime in a stable steady state (with approval

[29] If $R=0$ and $A_{i+1}=f\left(1/2\right)>1/2$, the phase diagram may have only one crossing point in case it starts above the 45-degree line, but since it starts above the 45-degree line and ends below it, this unique crossing-point must be a stable steady state.

[30] In the special case where $R=0$ and the phase diagram starts above the 45-degree line and has only one crossing point (which was shown to be a stable steady state), a decrease of $A_{i+1}=f\left(A_i\right)$ results as well in a decrease of A_{ss}.

A_{ss}), such that the size of the shock is larger than $|A_{ss} - A_{uss}|$ (when A_{uss} exists), would result in a revolution. A negative shock to the force of the regime reduces A_{i+1} (part (5) of Lemma 3), and in particular if the shock is such that $\bar{K}$ goes below $\bar{K}_{c3}$ A converges to zero over time and the regime completely falls (part (7) of Lemma 3). Finally, implementation of unpopular policies means that $|R|$ increases, and as a result the approval function (f)of the regime decreases (part (4) of Lemma 3), and in particular a revolution would start if the approval decreases sufficiently to eliminate the pre-existing stable steady state.

Part (3): (a) The fact that the whole population participates in the revolution follows from the fact that nobody in society fully follows the regime and this also implies (b) that the revolution wil be two-sided.

Part (4): Follows from part (3) of Proposition 1.

References

[1] Acemoglu, D., & Robinson, A.J., (2001). "A Theory of Political Transitions." *American Economic Review*, 91(4): 938-63.

[2] Al Jazeera (2011), "Timeline: Egypt's revolution A chronicle of the revolution that ended the three-decade-long presidency of Hosni Mubarak." February 14, 2011. http://www.aljazeera.com/news/middleeast/2011/01/201112515334871490.html. Accessed February 2017.

[3] Angeletos, G. M., Hellwig, C., & Pavan, A. (2007). "Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks," *Econometrica*, 75(3), 711-756.

[4] BBC (2013), "Profile: Egypt's Muslim Brotherhood", December 25 2013. http://www.bbc.com/news/world-middle-east-12313405. Accessed February 2017.

[5] Bernheim, D.B., (1994), "A Theory of Conformity", *Journal of Political Economy,* Vol. 102, No. 5, pp. 841-877.

[6] Breslauer, G.W., (2002). "*Gorbachev and Yeltsin as leaders*". Cambridge University Press.

[7] Brinton, C. (1938). "The Anatomy of Revolution". New York, NY, US: W W Norton & Co.

[8] Brown, A., (1997). "*The Gorbachev Factor*". OUP Oxford.

[9] Bueno De Mesquita, E. (2010). "Regime change and revolutionary entrepreneurs". *American Political Science Review*, 104(03), 446-466.

[10] Edmond, C. (2013), "Information Manipulation, Coordination, and Regime Change", *Review of Economic Studies*, Vol. 80, pp.1422–1458

[11] Esteban, J. (2001). "Collective action and the group size paradox." *American Political Science Association* Vol. 95, No. 03, pp. 663-672.

[12] Esteban, J., & Ray, D. (2001). "Social decision rules are not immune to conflict". *Economics of Governance*, 2(1), 59-67.

[13] Garner, P., (2001), Porfirio Díaz. New York: Pearson.

[14] Ghamari-Tabrizi, Behrouz. 2008. Islam and Dissent in Postrevolutionary Iran. New York: I.B. Tauris.

[15] Goldstone, J. A. (2001). "Toward a fourth generation of revolutionary theory". *Annual Review of Political Science*, 4, 139-187.

[16] Gorbachev, M. (1987), *"Perestroika. New thinking for our country and the world"*.

[17] Granovetter, M., (1978), "Threshold Models of Collective Behavior", *The American Journal of Sociology,* Vol. 83, No. 6, pp. 1420-1443.

[18] Kaniovski, Y. M., Kryazhimskii, A. V., & Young, H. P. (2000). "Adaptive dynamics in games played by heterogeneous populations". *Games and Economic Behavior*, 31(1), 50-96.

[19] Katz, F., (1998), The Life and Times of Pancho Villa. Stanford University Press.

[20] Kim, Q. Y. (1996). "From protest to change of regime: the 4–19 Revolt and the fall of the Rhee regime in South Korea". *Social Forces*, 74(4), 1179-1208.

[21] Kuran, T. (1989). "Sparks and prairie fires: A theory of unanticipated political revolution", *Public Choice*, 61(1), 41-74.

[22] Kuran, T., (1991), "Now out of Never, The element of surprise in the east European revolution of 1989", *World Politics*, Vol 44, No 1 pp. 7-48.

[23] Kuran, T., (1995), "The Inevitability of Future Revolutionary Surprises," *The American Journal of Sociology*, Vol. 100, No. 6, pp. 1528-1551.

[24] Kuran, T., & Sandholm, W. H. (2008). "Cultural integration and its discontents". *The Review of Economic Studies*, 75(1), 201-228.

[25] Lesch, A.M., 2011. "Egypt's spring: Causes of the revolution". *Middle East Policy*, 18(3), pp.35-48.

[26] Lohmann, S. (1994). "The dynamics of informational cascades". *World politics*, 47(1), 42-101.

[27] Manski, C.F.,Mayshar, J. (2003)"Private Incentives and Social Interactions: Fertility Puzzles in Israel," *Journal of the European Economic Association*, Vol. 1, No.1, pp. 181-211.

[28] Michaeli, M. & Spiro, D., (2015), "Norm conformity across societies," *J. of Public Economics*, Vol. 132, pp. 51-65.

[29] Milani, M. M. (1988). The making of Iran's Islamic revolution: from monarchy to Islamic republic. Boulder, CO: Westview Press.

[30] Moaddel, M. (1992). "Ideology as episodic discourse: the case of the Iranian revolution". American Sociological Review, 353-379.

[31] Naylor, R. (1989). "Strikes, free riders, and social customs". *The Quarterly Journal of Economics*, 104(4), 771-785.

[32] Oliver, P. E., & Marwell, G. (1988). "The Paradox of Group Size in Collective Action: A Theory of the Critical Mass". *II. American Sociological Review*, Vol. 53, No. 1, pp.1-8.

[33] Olson, M., (1971), *The Logic of Collective Action: Public Groups and the Theory of Groups*. Cambridge and London: Harvard University Press.

[34] Olsson-Yaouzis, N. (2012). "An evolutionary dynamic of revolutions". *Public Choice*, 151(3-4), 497-515.

[35] Pan, P. P. (2008). *Out of Mao's shadow: the struggle for the soul of a new China.* Simon and Schuster.

[36] Pfaff, S. (2006). *Exit-voice Dynamics and the Collapse of East Germany: the Crisis of Leninism and the Revolution of 1989.* Duke university Press.

[37] Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America.* Cambridge University Press.

[38] Razi, G. H. (1987). "The Nexus of Legitimacy and Performance: The Lessons of the Iranian Revolution". *Comparative Politics*, 453-469.

[39] Rubin, J. (2014). "Centralized institutions and cascades". *Journal of Comparative Economics.*Vol 42, Iss 2, pp. 340–357

[40] Sanderson, S.K., 2015. "*Revolutions: A worldwide introduction to political and social change*". Routledge.

[41] Shadmehr, M. (2015a). "Extremism in Revolutionary Movements". *Games and Economic Behavior*, 94, pp.97-121.

[42] Shadmehr, M. (2015b). "Ideology and the Iranian Revolution". mimeo, University of Miami.

[43] Tanter, R., & Midlarsky, M. (1967). A theory of revolution. Journal of Conflict Resolution, 11(3), 264-280.

[44] Tullock, G. (1971). "The paradox of revolution". *Public Choice*, 11(1), 89-99.

[45] Walder, A. G., & Xiaoxia, G. (1993). Workers in the Tiananmen protests: the politics of the Beijing Workers' Autonomous Federation. The Australian Journal of Chinese Affairs, 1-29.

[46] Young, H. P. (1993). The evolution of conventions. Econometrica: Journal of the Econometric Society, 57-84.

[47] Young, H. P. (2015). "The evolutions of social norms", *Annu. Rev. Econ.* 2015. 7:359–87

[48] Zhao, D. (2001). The power of Tiananmen: State-society relations and the 1989 Beijing student movement. University of Chicago Press.