The dynamics of revolutions*

Moti Michaeli†, Daniel Spiro‡

October 31, 2022

Abstract

This paper answers two questions: 1) Why do revolutions sometimes start with moderate opponents of the regime, like in the Egyptian Arab Spring in 2011 and in the USSR in 1989-1991? 2) Why does the implementation of popular policies, such as Perestroika, sometimes trigger a revolution? Existing theories invariably predict that none of this should happen. We show that answers can be provided by extending Kuran's (1989) dynamic model of mass protests by letting dissenters choose not just whether to dissent, but also how much. In the model, regimes may differ in how they sanction small vs. large dissent; and societies may differ in how individuals perceive the cost of deviating from their own ideological convictions. Such a generalization provides a typology of revolutions with predictions about who – moderates or extremists – are more likely to start a revolution, what ideology they will express and how this will dynamically change during a revolution. The model also provides predictions about which policies may trigger a revolution and which will consolidate the regime's strength. On question 1: Moderates are more likely to start a revolution if individuals are sensitive to even small deviations from their ideology. This sensitivity makes moderates, who slightly disagree with the regime, voice their criticism. Extremists, on the other hand, are silenced because expressing their extreme views bears heavy sanctioning. This further implies an answer to question 2: A popular policy may trigger a revolution, because, indirectly, it spurs more people to become "moderate" hence speak their minds.

Keywords: Revolutions; Mass protests; Ideology; Soviet collapse; Perestroika; Arab Spring.

*We wish to thank Elias Braunfels, Jean-Paul Carvalho, Sylvain Chassang, Henry Chen, Mona El-Sherif, Israel Eruchimovitch, Bård Harstad, Håvard Hegre, Yuval Heller, Moshe Kim, Carl-Henrik Knutsen, Arie Melnik, Anirban Mitra, Kalle Moene, Manuel Oechslin, Elena Paltseva, Paolo Piacquadio, Maria Petrova, Anja Prummer, Debraj Ray, Erik Snowberg, Kjetil Storesletten, Patrick Testa, Arthur van Benthem, Sareh Vosooghi, Yikai Wang, and seminar participants at George Washington University, Oslo Business School, University of Oslo, Tilburg University, Uppsala University, SITE, Bar Ilan, Kings College, EUI and the ASREC, IMEBESS and NCBEE conferences and three anonymous referees for valuable comments. We thank Handelsbankens forskningsstiftelser for research funding under grant P18-0142.

†Department of Economics, The University of Haifa. Contact: motimich@econ.haifa.ac.il.

‡Department of Economics, Uppsala University. Contact: daniel.spiro.ec@gmail.com.

1 Introduction

The purpose of this paper is to answer two questions: (1) Why do revolutions sometimes start with moderate opponents of the regime? (2) Why does the implementation of popular policies sometimes trigger a revolution?

Consider, e.g., the Arab Spring in Egypt in 2011. The protests were initiated by moderate liberals and moderate conservatives (Lesch, 2011), while those most critical to Mubarak's regime – the Muslim Brotherhood and the Salafists – were the *last* to join (BBC, 2013; Al Jazeera 2011). Similarly, many of the communist-regime collapses in Eastern Europe in 1989-1991 started with protests by moderates (or regime insiders) (Lohmann, 1994, Przeworski 1991, Breslauer 2002). Not least, the protests in the USSR itself, which were also started by moderates (led by Boris Yeltsin), followed popular reforms (Perestroika, see Brown 1997 and Lane and Ross 1994). That such policies could ignite a revolution came as a surprise to both experts and academics (as documented by, e.g., Kuran 1991 and Lipset and Bence 1994). Following this surprise, Przeworski (1991, p.1) wrote 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".

In contrast to these observations, *all* current models of revolutions predict that revolutions can only be started by the most extreme opponents of the regime and that, if anything, popular reforms should reduce the chances of a revolution.

We show in this paper that a first answer to the two questions above can be provided by analyzing not only whether somebody protests but also *how much* that individual protests. In order to highlight the driving forces, we build a very simple *dynamic* model of revolutions where individuals can choose *the extent* of their dissent. More precisely, we take the standard dynamic machinery developed by Granovetter (1978) and Kuran (1989), where the strength of a regime depends on how much support it gets and where the support in turn depends on the regime's strength.[1] To it, we add a choice structure allowing the individual protester to choose how much to dissent. This choice structure has previously been analyzed statically in Michaeli and Spiro (2015).

A typology with three distinct types of revolutions emerges from the model:

[1] A large number of other papers use a model like Kuran's (1989) one as a basic building block, and then enrich it, e.g., Naylor (1989), Bueno de Mesquita (2010), Edmond (2013), Rubin (2014), Passarelli and Tabellini (2017) and Dagaev et al. (2019).

1. **Extremists start the revolution**: Those who are the most critical of the regime initiate the revolution by expressing ideologically extreme critique, and later less critical individuals join the protests but dissent less than the initiators.

2. **Moderates start the revolution**: The revolution starts with moderates expressing moderate critique, and later less moderate individuals join and express ideologically extremer critique.

3. **All groups start the revolution:** All groups increase their dissent but the strongest opponents of the regime lead the revolution by expressing the harshest critique throughout the protests.

These revolutions differ with respect to who starts a revolution and how dissent evolves over time. In particular, type 2 corresponds to the dynamics of the Arab Spring in Egypt and the fall of Communism in the USSR.

The structure of the model is very simple. Society consists of citizens who differ in how misaligned their ideology is with the regime's policy. Each citizen in society considers what ideology to express in public. Expressing an ideology far from one's own is painful. But expressing an ideology far from the regime's policy (large dissent) is punished harder than expressing an ideology close to its policy. The general strength of punishment is decreasing in the overall dissent, namely the more citizens protest, and the further their expressed ideologies are from the regime's policy. This might lead to a cascade where the more people protest, the more is each individual inclined to protest as well. The increased protest and the parallel weakening of the regime are what we call a revolution.

There are four driving assumptions behind the typology above. A) The individuals, who differ in their ideology, face a non-binary choice of dissent. One can think of the dissent choice as determining the "ideology" a protester expresses on a left-to-right scale where the regime's ideology is located along that scale too. B) As a regime gets weaker, i.e., its ability to sanction dissent is lowered, each person becomes more inclined to protest. C) Not all regimes sanction dissent the same way. Some regimes sanction small dissent very heavily while large dissent is sanctioned only marginally more (a concave sanctioning); other regimes hardly sanction small dissent, while sanctioning is ramped up considerably for large dissent (a convex sanctioning).[2] D) Societies differ "culturally"

[2]Previous research and anecdotal evidence indeed suggest regimes differ in how they sanction dissent. In Iran in 1979 there are indications of a concave sanctioning as the Shah suppressed all manifestations of political dissent, moderate and extreme alike (Milani 1988, p122) and was treating the

in how citizens view a misrepresentation of their true views. In some societies individuals consider any small misrepresentation of one's true ideology as very costly, while larger misrepresentation is only marginally more costly (a concave ideological cost); in other societies individuals consider small misrepresentations of the self rather costless, but large misrepresentation is very costly (convex ideological cost).[3]

We show that the evolution of *dissent* depends on the sanctioning structure the regime is using. Roughly speaking, 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. Thus, if dissenting, an individual may as well express his true views. This implies that an individual or a group that increases their dissent will do so non-marginally – going from silence to considerable dissent. However, if the sanctioning cost is convex, then small dissent is not so costly while large dissent is very costly. Hence, under such sanctioning, revolutions start with small dissent and, as the regime gets gradually weaker during the revolution, the dissent gradually increases – the freedom of speech is pushed further.

The evolution of *participation* (whether moderates, extremists, or both groups start protesting) depends on the cost of deviating from one's ideological bliss point. In a society that is characterized by individuals with a convex ideological 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. This means revolutions will be started by individuals whose private views are very far from the regime's policy. This aligns with previous models. In contrast, in a society that is characterized by individuals with a concave ideological cost, individuals will find it very costly to deviate even a little from their bliss points, but deviating more will be only marginally more costly.[4] Hence, if they do deviate, they might as well align with the regime, for instance by remaining silent. Those who will find

entire opposition movement as a homogeneous entity (ibid, p. 197). In contrast, in the Tiananmen-square protests in 1989, accounts suggest sanctioning was convex as the initial (moderate) protests were largely tolerated and sanctioning was ramped up only when the dissent escalated (Walder and Xiaoxia, 1993; Zhao 2001; Pan, 2008). More broadly, other research suggests that deviations from norms (Hermann et al, 2008) and religious conduct (Michaeli and Spiro, 2015) differ across societies.

[3]Research indeed suggests there are differences in this curvature, depending on the society and the ideological dimension at hand. For instance, a convex cost has been found to fit patterns of lying (e.g., Fischbacher & Follmi-Heusi 2013, Abeler et al. 2019; but see contrary arguments for a concave/fixed cost in Kajackaite and Gneezy 2017, Gneezy et al. 2018), whereas a concave cost was found to fit voting (Kendall et al. 2015) and judicial decision making (Chen et al. 2019).

[4]These concave costs can be interpreted as risk-loving ideological or religious preferences – individuals would be willing to take risks in order to be able to adhere precisely to their own ideology, morals or religion.

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. Consequently, if a revolution starts, it will be started by *moderates*. This pattern can explain the unfolding of the revolutions in Egypt (2011) and the USSR (1989-1991) and thus answer the first research question. In this scenario, great misalignment with the regime's policy silences an individual. Hence, if the regime's policy is misaligned with large parts of society there will be less protest. The answer to the second research question then follows from the observation that implementation of popular policies makes more people inclined to speak their minds, which can ultimately start a revolution.

To highlight the driving mechanisms, our theory abstracts from many elements present in other models. This includes collective action (Olson, 1971; Tullock, 1971; Chwe, 1999; Esteban and Ray, 2001), economic forces (Davies, 1962; Hegre and Sambanis, 2006; Knutsen, 2014), non-violent strategies (Desai et al. 2019) and resource-mobilization ability (see Jenkins, 1983 and Edwards and Gillham 2013 for summaries). In the basic model we also abstract from strategic considerations. This does not drive our results. In an extension we allow the regime to make strategic choices. More precisely, the regime optimally chooses its sanctioning structure to minimize dissent. We show that our typology remains valid and, most importantly for our purpose, that revolutions can be started by moderates even when the regime chooses its sanctioning optimally. We also show that regimes will tend to choose a more convex sanctioning structure the more extremists there are in society.[5]

While no previous research provides answers to our research questions, three papers are closely related to ours from a modeling-perspective. Chen and Suen (2020), like us, present a protest model that is perhaps best interpreted in terms of ideology as a driver of political unrest. Their agents, like ours, are infinitesimal and non-strategic. In particular, Chen and Suen (2020) set out to explain why the agenda of some protest leaders is seemingly so ideologically extreme that it hurts their own chances of success (it builds on a signaling game).[6] Thus, while they do let the ideological choice of protest leaders be chosen from a continuum, they do not explain why some revolutions are initiated by moderates. Furthermore, their model does not have dynamics thus cannot

[5]We have also analyzed a strategic (and dynamic) choice of policy by the regime, and solved a version of our model with a small number of strategic revolutionary factions. Our main results remain the same, in particular the prediction that moderates may be the ones initiating the revolution. We do not include these extensions here, to keep the paper at reasonable length.

[6]See Enikolopov et al. (2020) for another recent signaling model.

explain why individual positions would change over time.

Shadmehr (2015), like us, differentiates between different extents of revolutionary actions and, as such, is able to analyze how extreme the revolutionary agenda will be. He also studies how a regime's sanctioning structure (its curvature) affects participants' choices, which is important in our paper too. However, his model has no dynamics – the revolutionary agenda is assumed to be constant throughout the revolution – and predicts that the most extreme factions will under all circumstances be part of the revolution, hence it does not provide answers to our research questions.

Finally, as mentioned, we use the same individual choice structure as in Michaeli and Spiro (2015), which analyzes whether individuals will make large or small deviations from a social norm. The model there is not dynamic and the strength of the norm is exogenous. Thus, being static, that model cannot explain how participation and the extent of dissent will change during a revolution. To that model we add here a dynamic structure, an endogenous regime strength and, in an extension, an endogenous choice of sanctioning structure by the regime. In the literature on social norms, other related models such as Bernheim (1994) and Manski and Mayshar (2003) are, again, static. Dynamic models of social norms we are aware of are Kuran and Sandholm (2008) and Michaeli and Spiro (2017). In the latter, there are also results pertaining to the curvature of individual tastes, but, importantly, an individual does not relate to the sanctioning of a single regime or norm but to the whole distribution of other individuals, and this has a significant effect on the results and on the questions that the model can answer.

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 mass 1 and of a political regime. The regime has a policy $R \in [-1, 1]$. Focusing on revolutions and mass protests against a given regime, just like Kuran (1989), we let R be exogenous (capturing the regime's ideology). Each individual takes a "political" stance (action) $x \in \mathbb{R}$. The distribution of stances is denoted by X. The stance x can be interpreted as the expression of a political opinion on a left-to-right scale where $x = R$ means the individual expresses agreement with the regime (or, alternatively, stays silent) and $x \neq R$ can be interpreted as the individual criticizing or protesting against the regime. The regime sanctions dissent, i.e., stances that deviate from its policy ($x \neq R$). As a guiding principle, we assume that sanctioning is increasing in the distance between x and R, i.e.,

harsher critique of the regime is sanctioned more heavily:

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

The severity of sanctioning, to which we also refer as the *strength* of the regime, is represented by K in (1). It is endogenously determined by

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

where $\bar{K}$ is a parameter capturing the *force* of the regime and $A \in [0, 1]$ is the *approval* of the regime, an endogenous variable which in itself is determined by the overall dissent in society. $\bar{K}$ could represent, e.g., the per capita law-enforcement forces used by the regime to sanction dissent.[7] We make the minimal necessary assumption regarding the regime's approval: A is smaller the larger is the number of dissenting individuals and the more dissenting each public stance is.[8] Hence, just like in Kuran (1989), the severity of the regime's sanctioning (K) decreases in total dissent. This captures, for instance, the notion that increased dissent makes it harder for the regime troops to catch any particular dissident, or makes the troops less assertive when sanctioning dissent.

The β parameter captures the curvature of the sanctioning system. We allow β to take any value larger than zero since previous research shows that this curvature indeed differs across regimes and societies (see introduction). This generalization is important for the analysis. A regime with a large β (> 1) uses convex sanctioning and hence is tolerant towards dissent as long as it is not too extreme. 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.

Having defined the external cost of taking a stance x, we move on to define the internal cost: the equivalent of Kuran's (1989, p.47) "psychological [cost that an agent] suffers for compromising his integrity". 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*,

[7]The parameter $\bar{K}$ could be incorporated into the approval A. However, we deliberately separate K to its two ingredients, so that the approval is normalized to 1 while $\bar{K}$ has the interpretation of capturing the regime's force (on which we later do comparative statics).

[8]This is the equivalent of Kuran's (1989) assumption. More precisely, A has the property that for any non-zero mass of individuals, if these individuals strictly increase their dissent (while holding the dissent of the rest fixed) then A strictly decreases (when starting from $A > 0$). Note that, to obtain our results, we do not need to make further assumptions about A, i.e., our results hold for any specification of A having the aforementioned property. E.g., we do not need to make specific assumptions on what happens to A if one individual increases dissent while another individual decreases it.

with $g(t)$ denoting the probability-density function of types, which is assumed to be continuous. When expressing a stance x, the individual bears a cost for deviating from the bliss point:

$$D(x,t) = |t-x|^{\alpha} \ , \ \alpha > 0. \tag{2}$$

D can be interpreted as *discomfort* from expressing a political opinion not in line with a person's conviction. That is, following for instance Kuran (1989), Goldstone (2001) and Rubin (2014), individuals can be viewed as driven by a will for self expression. Thus, $|t-R|$ captures how deviant the individual's bliss point is with respect to the regime's policy, while $|t-x|$ captures how much the individual restrains dissent. For short, we 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.[9] The generality in allowing the parameter α to take any value larger than zero is based on previous research (see introduction) and is important for our analysis. This parameter captures how an individual perceives small versus large deviations from the bliss point. It can be viewed as a cultural, societal or religious attribute and is assumed fixed across the population. With a large $\alpha(>1)$, an individual is insensitive to small deviations from the bliss point, while large deviations are very costly. This further implies extremists will find it very costly to fully follow the regime while for moderates this will be nearly costless. With a small $\alpha(<1)$, an individual perceives even small deviations as very costly, but hardly distinguishes between small and large deviations. This further implies it will be almost equally costly to follow the regime for extremists and moderates alike.

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 the privately held opinion. That is, the individual chooses x to minimize

$$L(x;t,R,K(X)) = D(x,t) + S(x,R,K(X)). \tag{3}$$

It is immediate from this choice problem that each individual t 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 $K(X)$ and hence indirectly on the stances taken by all individuals in society.

[9]We wish to emphasize that, in our model, the difference between moderates and extremists is along ideological lines only, they are otherwise identical.

Having outlined the properties of the static model, we can now define what a Nash equilibrium entails in our setting: it is a mapping $t \to x$ where each individual is best responding (minimizing (3)) given the actions of the other agents. The stance that minimizes L for an individual with opinion t is denoted by $x^*(t)$, and the distribution of all such chosen stances by the population is denoted by X^*. Just like in any game with complementarities between the actions of agents there may exist multiple equilibria in our static game. For example, an equilibrium with a strong regime and small dissent may exist and an equilibrium where the regime is weak and dissent is large may exist too.

We add now a simple dynamic structure to the model in order to analyze the movement from a state where the regime is strong to one where it is weaker (a *revolution* – see formal definition below). We use the same dynamics implicitly used in Kuran's (1989) model and which are standard in games with large populations (e.g., Young 1993; Bala and Goyal 2001; Young 2015; Bursztyn et al., 2019). In these dynamics, agents are essentially best responding in period $i+1$ to the actions taken in period i:

$$x^*_{i+1}(t, R, K(X^*_i)) = \arg\min_{x_{i+1}}\{L(x_{i+1}; t, R, K(X^*_i))\}, \tag{4}$$

which we for short denote by $x^*_{i+1}(t)$. We wish to emphasize that these adaptive dynamics do not drive our results (see footnote 5). Their merit is in creating smooth dynamics instead of unrealistic jumps of the whole population between steady states and in allowing analysis of shocks and convergence. In practice we assume an individual's "personal influence on the selection of the social order is negligible" (Kuran 1989, p.47).

Equation (4) implies that the regime strength (K) that affects stances in period $i+1$ is determined by the stances taken in period i. This could capture, e.g., the assertion of the regime's troops at day $i+1$ of a revolution after observing the dissent of the population on the previous day. This dynamic structure implies that the distribution of stances X^*_{i+1} is a function of X^*_i and that

$$A_{i+1} = f(A_i) \tag{5}$$

where f describes the dynamics of approval between periods. A steady state is reached when $x^*_{i+1}(t) = x^*_i(t)\ \forall t$, i.e., when each person's best response in period $i+1$ equals his action in period i. Hence, a steady state also constitutes a Nash Equilibrium. In such a situation, $f(A_i) = A_i$. We consider a steady state to be stable, with its approval denoted

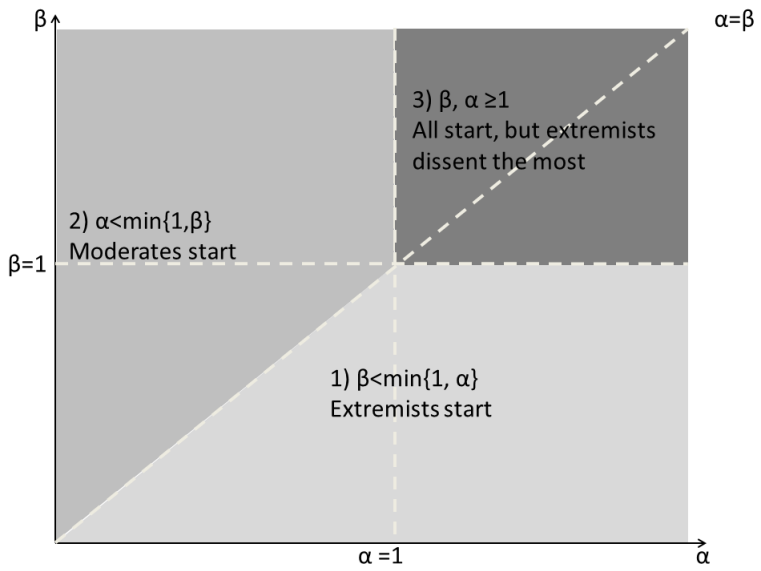

Figure 1: Parameter space of the different classes of revolutions.

by A_{ss}, if there is convergence back to it following a small perturbation to $A_i = A_{ss}$. 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 $A \in (A_{uss}, A_{ss})$. We define a revolution as follows.

Definition 1. A revolution is a sequence of time periods in which dissent in the population increases hence approval decreases.

3 Main result

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 a revolution. 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.[10]

Proposition 1. *There are three distinct and exhaustive classes of revolutions:*

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

1. *(**Extremists start the revolution**) When $\beta < \min\{1, \alpha\}$: 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 start the revolution**) When $\alpha < \min\{1, \beta\}$: Initially the dissenters are moderates, and later in the revolution types who are more extreme join and dissent more than the initial moderates.*

3. *(**All groups start the revolution**) When $\beta \geq 1$ and $\alpha \geq 1$ (and at least one inequality is strict): All groups increase their dissent throughout the revolution, where the most dissenting types are always the extremists.*

Proof: See Appendix A.2.

These three classes of revolutions fundamentally differ in how participation and statements evolve during the revolution. Part 2 provides an answer to our first research question. We defer explaining the full intuition of Proposition 1 to the upcoming three sections, where we analyze each class of revolutions separately and describe further results.

The following corollary highlights the overall pattern and follows directly from the proposition.

Corollary 1. *For any given β, extremists take part in starting the revolution if and only if α is sufficiently large.*

Figure 1 is instructive for highlighting the overarching pattern given by the corollary. Moving along the horizontal axis, the further to the right we are (large α), the more likely it is that extremists take part in starting the revolution. That is, whether individuals distinguish between small and large deviations from their bliss points determines who the *initiators of the revolution* will be – moderates or extremists (or both). In particular, if individuals are largely insensitive to small deviations from their bliss points but perceive large deviations as very costly (large α), then extremists are more likely to take part in starting a revolution since they perceive it as very costly to keep silent. On the other hand, if individuals are very sensitive to small deviations from their bliss points, but do not distinguish much between small and large deviations (small α) then instead moderates are more likely to start a revolution because they are sanctioned less than extremists when speaking their minds.

Previous research suggests there is heterogeneity in the curvature of ideological cost, with differences across societies and across ideological dimensions (Kendall et al., 2015;

Chen et al, 2020; Abeler et al. 2019). Hence, the corollary provides a prediction for how the underlying preference structure affects whether moderates or extremists are more likely to start a revolution. This prediction is in principle testable with the caveat that, to our knowledge, the necessary data does not currently exist. While the NAVCO data set (see Chenoweth et al, 2017) has some information on the ideology of insurgent groups, a systematic assessment of the ideological costs across societies does not exist. However, an emerging body of research has started developing ways of measuring the curvature of preferences (see e.g. Baranski et al. 2020). If such methods become feasible to conduct across societies and possibly across time, then our theory can be used not only to rationalize the puzzling events presented but can also be tested as per Corollary 1.[11]

The results stated in Proposition 1 hold for any continuous distribution of types $g(t)$ and without setting any restriction on the functional form of the approval function A.[12] While intuition suggests that this is true also for most of the other results in the paper, we cannot show it formally, as it becomes extremely complicated to prove them without choosing a more explicit functional form of the distribution of types and of A. We will thus assume in the next three sections that $t \sim U(-1,1)$ and that the approval of the regime is linear in the deviations from it. Let

$$\Psi\left(x_i; R, A_i\right) \equiv \int_{-1}^{1} \left|x_i\left(t\right) - R\right| dt. \tag{6}$$

Then

$$A = \max\left\{0, 1 - m\Psi\left(x_i; R, A_i\right)\right\}. \tag{7}$$

This is a specific linear approval function A – the approval linearly decreases in the sum of absolute deviations from R. $A = 1$ if nobody dissents ($x\left(t\right) = R\ \forall t$). Since $E\left[t\right] = 0$ we say that the regime is *biased* with respect to people's preferences if $R \neq 0$

[11] Baranski et al. (2020) use an incentivized scheme for eliciting the curvature of the cost of ideological deviations by asking subjects for their *minimum acceptable amounts* (MAAs) – the amounts to be given to them so that they would be okay with the experimenter donating $100 to interest groups with clearly-identified ideological attributes (e.g., the National Rifle Association on one end of the spectrum and the Coalition to Stop Gun Violence on the other end, where attitudes towards organizations in between these two extremes serve for identifying the curvature of the ideological cost). Other studies, such as Inglehart and Welzel (2005) and Dahlum and Knutsen (2017) use cultural traits for studying institutions and conflict. In these studies they focus on self-expression and they use the World Value Survey. By adding questions that use the MAA-elicitation technique to such a survey, it should in principle be possible to get data on the cross-country differences with respect to how picky people are about deviating from their ideological bliss point.

[12] In fact they can be derived also with more general functional forms for S and D.

and, accordingly, we refer to $|R - E[t]|$ as the bias of the regime. We normalize m so that a non-biased regime has zero approval precisely when all types speak their minds ($x(t) = t \; \forall t$). The normalization is therefore $m = 1/\left(2\int_0^1 tdt\right) = 1$. This normalization is largely without consequence apart from implying that even a central regime loses all of its strength (A becomes 0) when all speak their minds (which would not be true for $m < 1$).[13] The state where all speak their minds can be interpreted as a state with no regime or with a regime that has no control over the population or with a regime that does not sanction dissent – in practice freedom of speech prevails.

The following concepts are useful for our further analysis.

Definition 2. A successful revolution is one where $A = 0$ at its end and a failed revolution is one where $A > 0$ at its end.

By Definition 2, each successful revolution ends in a freedom of speech (since $A = 0$ implies $K = 0$). We refer to the triggers of revolutions as *catalytic events*: changes or shocks that either imply that a previously stable steady state ceases to exist, or decrease the approval to a point where it will, endogenously, decrease further.

4 Extremists start the revolution

We start by considering the case $\beta < \min\{\alpha, 1\}$ where, by Proposition 1, any potential revolution entails extremists starting it. This case can be further divided into two subcases: $\beta < \alpha \leq 1$ and $\beta \leq 1 < \alpha$. 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 $\beta < \alpha \leq 1$.[14]

The full properties of this revolution, which largely align with most previous models, are outlined 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 the implementation of a new policy only if it is unpopular.*

[13] The normalization 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).

[14] See sections A.1.2 and A.2.1 in the appendix for a treatment of the other subcase ($\beta \leq 1 < \alpha$).

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.*[15]

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

Proof. See Appendix A.3. ■

The key property of this case is that β is 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 naturally induce an individual to either fully follow the regime or, if not, the individual might as well dissent as much as he likes.[16] Then, since extremists perceive the highest discomfort (D) when following the regime, they will be the ones who may dissent – and, given their extreme views, dissent extremely if they do – while moderates will be silent (point 3 and 4 of the proposition).

A technical remark is now in place. The above logic implies that each type has a unique best response $x^*_{i+1}(t)$ for any given level of regime approval A_i.[17] Then, since X^*_i determines the new approval A_{i+1}, we have that each A_i maps into a unique A_{i+1}. This is convenient since it implies that the dynamic properties of the model can be collapsed into one static as per the endogenous function $A_{i+1} = f(A_i)$. Along with the 45-degree line, where $A_{i+1} = A_i$, the function $A_{i+1} = f(A_i)$ creates a phase diagram (see Figure 2). The properties of $f(A_i)$ (whose analysis is the main focus of the proof) determine the existence and number of steady states, whether they are stable and in which direction convergence goes. Thus, together with the best response of all individuals ($X^*_i(A_i)$), the function $f(A_i)$ determines the full dynamic properties of the model. Each parameter setting (β and α) will have its own pattern of $X^*_i(A_i)$ and $f(A_i)$, and this pattern is what distinguishes the three types of revolutions.

The properties of $x^*_i(t)$ in the case analyzed here – extremists dissent, moderates stay silent – imply that a large misalignment between an individual's ideology and the

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

[16] By differentiating L (in (3)) twice with respect to x_i it is immediate that when $\beta < \alpha \leq 1$ the second-order condition is not fulfilled, implying that an individual will choose either $x_i(t) = R$ or $x_i(t) = t$.

[17] Bar one cutoff type who is indifferent and whose weight is zero.

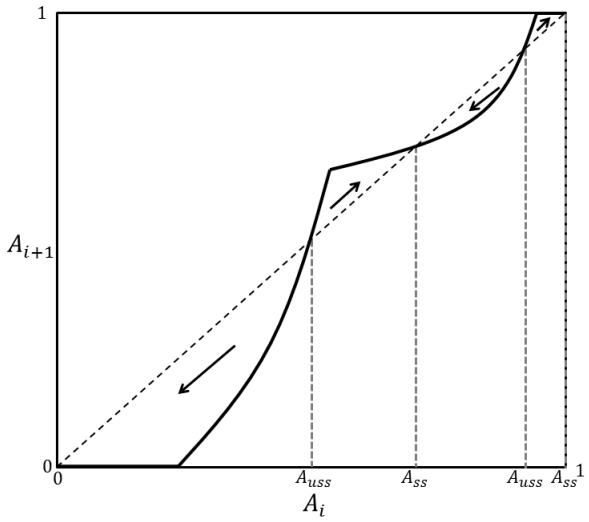

Figure 2: $\beta < \alpha \leq 1$ and a moderately biased regime, $|R| \leq 0.5$. The solid line depicts the intertemporal-dynamics function $A_{i+1} = f(A_i)$ and the dashed line depicts the 45-degree line where $A_{i+1} = A_i$. The vertical lines depict the stable (A_{ss}) and the unstable (A_{uss}) steady states. Note that when the regime is very biased, $0.5 < |R| \leq 1$, the phase diagram will not contain the left convex part.

regime's policy is what triggers dissent. Consequently, the more biased the regime's policy is, the more force it needs to employ (point 1). This further implies that an increase in the misalignment between the regime and the population as a whole may trigger a revolution (point 2).[18] Such misalignment may be the result of either the regime implementing unpopular policies or a shift in the preferences of the population (the t-distribution) away from the regime. Figure 3 illustrates what a revolution would entail in the latter case.[19] In the top schedule we are in a steady state where the regime is non-biased ($R = 0$) and is sufficiently strong to induce no dissent. In the figure this can be seen by the distribution of statements being completely centered on R. In the next schedule (representing a later time period), the preferences of the population have shifted to the right to the extent that the increased misalignment between the population and the regime triggers some dissent on the far right. This reduced approval weakens the regime and this in turn, in the third schedule, enables more moderate types on the right

[18] In the phase diagram in Figure 2, increased bias has the effect of lowering the approval function $f(A_i)$ thus destabilizing steady states. Other potential catalytic events, which in fact apply to all the three classes of revolutions, are shocks to either the regime's force or to its approval. Such shocks, if large enough, imply convergence to a new, lower level of approval (potentially even $A = 0$, implying a collapse of the regime).

[19] An example for a revolution that seems to fit this pattern is the Iranian Islamic Revolution against the Shah in 1979. The extremists that initiated this revolution were Khomeini and his followers, who demanded right from the start a radical change of the regime in Iran into a religious Islamic state. The increased misalignment between the Shah's secular policies and the increasingly religious sentiments in society is documented in Moaddel (1992).

to speak their minds too. The increased dissent weakens the regime's sanctioning further, enabling even more moderate individuals to dissent, eventually implying people on the left start dissenting too (bottom schedule). Hence, overall, we get that the revolution is initiated by extremists dissenting extremely, while more moderate individuals join the revolution later and dissent less.

An interesting aspect is that this revolution is one-sided initially (dissent comes only from one side of the political spectrum). Since new recruits come only from one side, the momentum of the revolution will be weak initially, implying further that during this phase the revolution may fail (such a failed revolution is represented by convergence to the second stable steady state from the right in Figure 2). However, once the revolution becomes two-sided (A_i is to the left of the middle kink in the phase diagram), the momentum picks up and the regime is bound to collapse – the approval goes to zero. If the revolution fails, the regime becomes what one might call "semi-democratic", where some individuals speak their minds – in practice some freedom of speech exists. This regime will be weaker and will more likely collapse should a subsequent revolution break, in line with the empirical observation that semi-democracies are the least stable regimes while pure democracies and pure autocracies are the most stable (Gurr 1974; Gates et al. 2006; Knutsen and Nygård 2015).

5 Moderates start the revolution

We move now to the most interesting case, the one providing an answer for our two research questions: 1) why revolutions may be initiated by moderates and 2) sometimes triggered by popular policies. Here $\alpha < \min\{\beta, 1\}$. As Proposition 1 states, this case entails that each revolution will be started by moderates, while extremer individuals join only later and dissent more than the initial moderate dissidents. 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 $\alpha < \beta \leq 1$.[20]

The full properties of this revolution are outlined 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.*

[20]See sections A.1.3 and A.2.1 in the appendix for a treatment of the other subcase ($\alpha < 1 < \beta$).

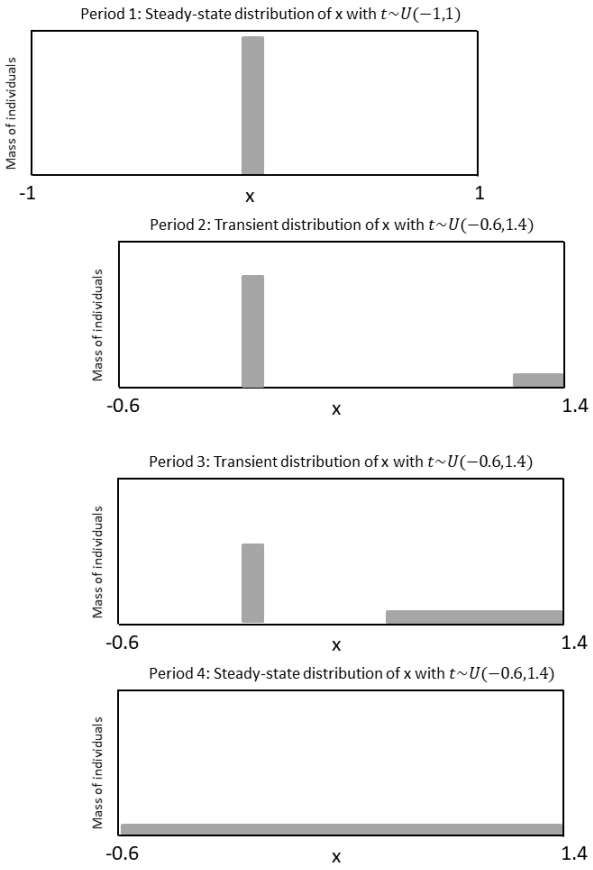

Figure 3: Distribution of stances over time in a stylized case of a revolution starting with extremists dissenting extremely ($\beta < \alpha \leq 1$). $R = 0$ and fixed while the distribution of types changes.

2. ***Catalytic events****: A revolution may start following the implementation of a new policy only if it is popular.*

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

   (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.*

Proof. See Appendix A.4. ■

To understand the intuition behind these results note that the key property of this case is that α is relatively small (and in particular smaller than β). It is best illustrated by considering α close to zero, implying D is a step function. Then an individual will perceive a high cost from even a small deviation from the bliss point, but will not distinguish between small and large deviations. This implies the agent will either speak his mind or, if this is too difficult given the sanctioning, be willing to go a long way to avoid punishment by the regime – there is no point in compromising on a stance in between the bliss point and the regime when D is as large when doing so as when staying silent. Then, since an extremist who speaks his mind is sanctioned heavily (large $|t - R|$ hence large S), the extremist will submit to the pressure and follow the regime. In a sense, the extremists are giving up on expressing their ideology if they cannot express it exactly as they wish, thus they stay silent. Meanwhile, for a moderate it will be equally costly to follow the regime as it is for the extremist (since D is a step function). But, compared to the extremist, the moderate will face lower sanctioning when speaking his mind, hence will prefer to do so over bearing the large personal discomfort of deviating from the bliss point. Put together, this implies that, if a revolution is started, moderates will be the first out dissenting. Speaking their minds, these moderates pose only mild critique of the regime. When the moderate individuals start dissenting, the approval (and thus strength) of the regime falls, which enables less moderate types to speak their minds too. This further weakens the regime's punishment, enabling extremer individuals to express their extremer views as well and so on. Thus, in contrast to the previous case and to all current models, the revolution here is started by moderates (expressing mild critique) while extremists join only later (points 3 and 4 of the proposition). This provides an answer to our first research question.

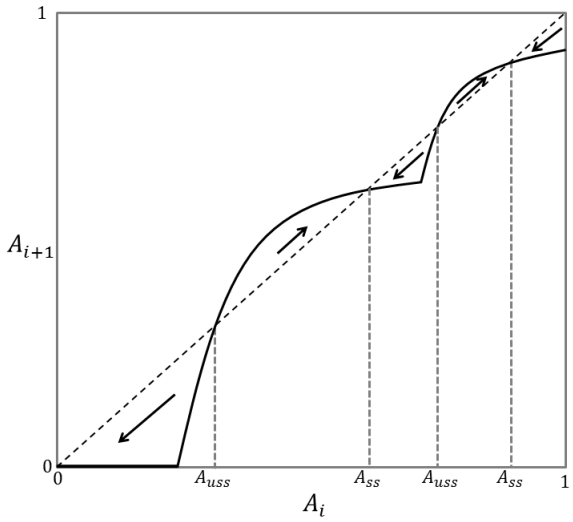

Figure 4: Stylized phase diagram for the case $\alpha < \beta \leq 1$ and a biased regime. The solid line depicts the intertemporal-dynamics function $A_{i+1} = f(A_i)$ and the dashed line depicts the 45-degree line where $A_{i+1} = A_i$. The vertical lines depict the stable (A_{ss}) and unstable (A_{uss}) steady states.

The phase diagram in Figure 4 illustrates the approval function, steady states and convergence zones to the stable ones (the full properties are outlined in Appendix Section A.4).

The above description implies that great misalignment between an individual's ideology and the regime's policy will induce the individual to *stay silent*. This has important implications for the stability of regimes and for what may trigger a revolution. A biased regime ($R \neq 0$) can employ less force yet remain stable (point 1): since the individuals whose private opinion is far from the regime's policy stay silent, a biased regime, whose policy is ideologically far from the opinions of many in society, will have more approval than a central regime with the same force. In Figure 5 this is illustrated on the left side (Case 1). Here, an increase in the misalignment between the regime's policy and people's preferences (when going from the first to the second schedule), due to a shift in the population's preferences away from the regime, induces less dissent.

The effect of implementing popular policies is the opposite. The regime policy then aligns with the views of more people and, since those who only slightly disagree with the regime are the ones dissenting, there will be more dissidents speaking their minds. Hence, a popular policy decreases the regime's approval and its implementation may ultimately be the catalytic event that starts a revolution (point 2). That is, unlike with revolutions started by extremists, here popular policies can trigger a revolution. This provides an answer to our second research question.

An equivalent logic applies when people's preferences move to align more with the

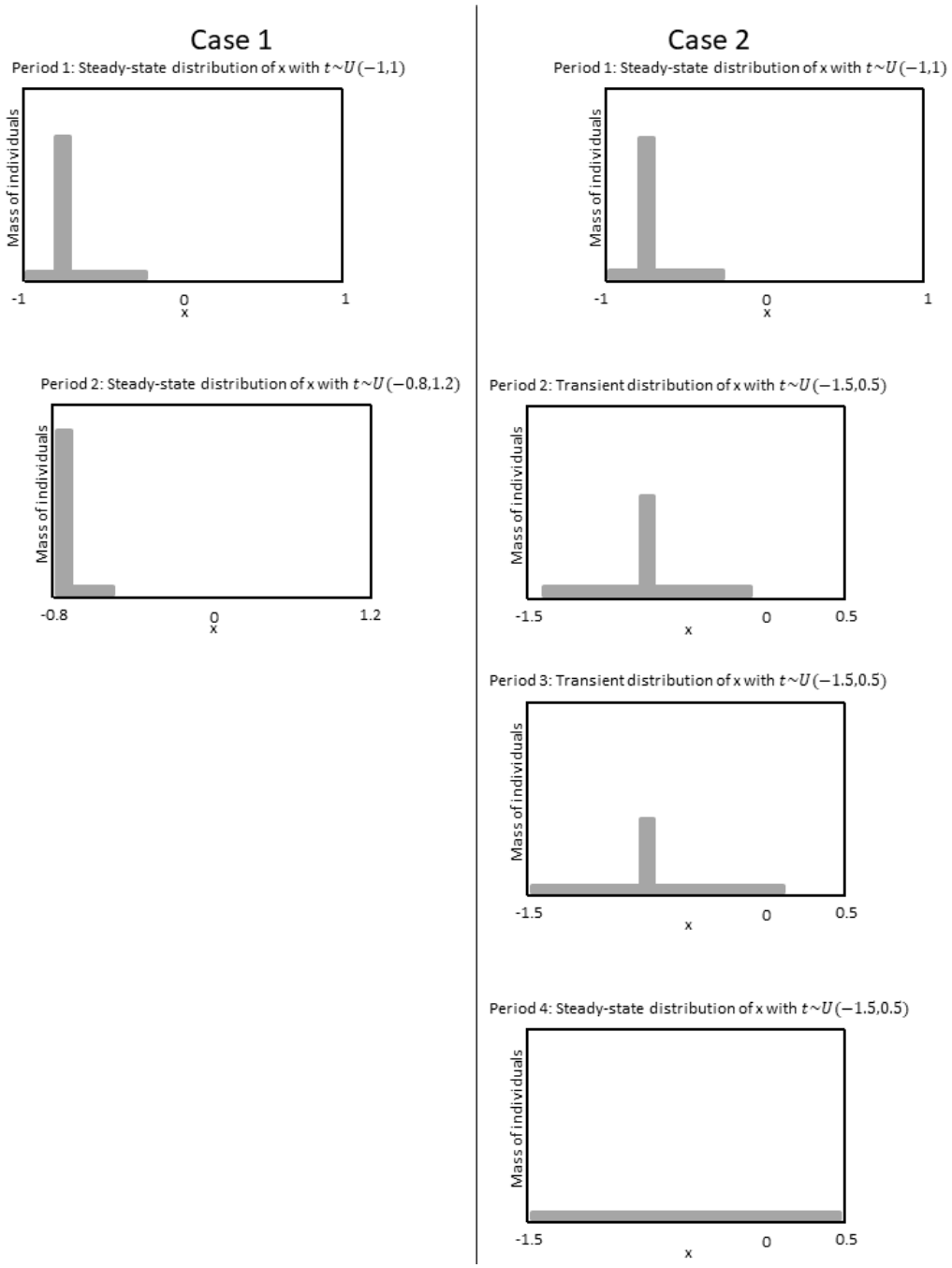

Figure 5: Distribution of stances over time when $\alpha < \beta \leq 1$. In both cases $R = -0.8$ and fixed while the distribution of types changes. Case 1: the shift of private preferences to the left does not trigger a revolution. Case 2: the shift of private preferences to the right does trigger a revolution starting with moderates dissenting moderately.

regime, as illustrated on the right part (Case 2) of Figure 5. In the top schedule we start in a steady state where the regime is very biased to the left. In the second schedule, the population's preferences have shifted to the left, i.e., they have become more aligned with the regime. This induces the "new moderate left" to dissent, which reduces the regime's approval thus its strength and enables extremer individuals both on the left and on the right to join (third schedule). This reduces the approval further, which enables also the most extreme rightists to finally speak their minds. This way, what started as a leftist revolution, following a leftward movement of the population's sentiments, ends up being a rightist revolution, where the center of expressed opinions is eventually to the right of the regime that collapsed, revealing that society was all along more rightist than the regime.

The fact that the revolution will be two-sided from the beginning implies that its momentum will be strong initially – new dissenters will join from both the left and the right. Eventually, however, as depicted in the figure, the potential new recruits are exhausted on the left, and the momentum will be reduced. At these later stages the revolution may fail (in Figure 4 this is represented by convergence to the second stable steady state from the right; see a formal explanation in Appendix A.4). Thus, unlike the revolution started by extremists, here the revolution is strong initially and weak later.

5.1 Historic examples

The revolutionary pattern just described is in stark contrast to what all models of revolutions predict. But the dynamics described provide a theoretically-consistent explanation for an important class of revolutions and mass protests. For example, it aligns with the collapse of the USSR (and some of the protest movements in Eastern Europe) and to the recent Arab-Spring revolution in Egypt. We will briefly describe these revolutions and protests through the lens of our theory. In Appendix B we discuss some alternative mechanisms and their ability to explain these events.

5.1.1 The collapse of the USSR in 1989-1991

The fall of Communism in the USSR fits our description both in being triggered by Gorbachev's implementation of a popular policy (Perestroika, to be discussed shortly) and by the evolution of participation from moderates to extremists. The first to protest was indeed a party insider – Boris Yeltsin – who at various meetings in 1986-1988 openly criticized Gorbachev and his government for the reforms not being sufficiently

far reaching (Breslauer, 2002 p.130-132).[21] His dissent spread within the party to other, more liberal (i.e., extremer) members eventually joining Yeltsin in forming the inter-regional deputies group in 1989 (Lane and Ross, 1994). Yeltsin's insubordination enabled the formation of dissident groupings also outside the communist party in 1990.[22] The protests spread, not least with the help of these groups (Urban and Gelman 1997; Brudny 1993), to broader parts of society. They evolved to mass protests and rallies in 1990-91 demanding democratic reforms and economic liberalization far beyond Perestroika.

Beyond the progress of participation and statements from moderate to extreme, another important feature of this class of revolutions in our model is that the undermining of the regime is initiated by individuals from *both* sides of the political spectrum. 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. Indeed, this was the case in the USSR. The dissent against Gorbachev was two-sided early on (Sanderson, 2015, p.126). Apart from "liberal" Yeltsin, hard-line communists within the party criticized Gorbachev's liberalization reforms (but for being too far-reaching) and later even tried to overturn them by staging a coup (Lane and Ross, 1994). In sum, the dissent against Gorbachev increased from all directions and his total support fell dramatically.[23] Since the population in the USSR had a more liberal and more democratic inclination than the communist party (Gibson, 1997) this was the main direction the protests took.

The further puzzle is, of course, that 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 as a form of popular policy – in the hope to revitalize and modernize the Soviet Union –

[21] Also in other Eastern-European countries the initial protesters were moderates or even party insiders. For example, in Poland and Hungary, moderate dissidents instigated liberal reforms and made demands for free elections (Pfaff, 2006, p.1). Hungarian communist-party leader Karoly Grosz was quoted saying that "the party was shattered not by its opponent but – paradoxically – from within" (Przeworski 1991, p. 56).

[22] For example, the Democratic Russia Election Bloc and the Democratic Russia Movement were formed by grassroots, pro-democracy and anti-communist organizations and generally politically-unaffiliated persons. These groups had a liberal and democratic political agenda that was extremer than that of Yeltsin, with the main goals of removing the Communist Party from power, transforming the economy to be market based and ridding society of the communist legacy (Brudny, 1993). At various stages of the revolution some of these groups supported Yeltsin. However, the relationship between Yeltsin and these groups was never frictionless and collapsed when Yeltsin disagreed with the radical agenda these groups pursued (Brudny 1993).

[23] This is documented by voting records within the party (see Figures 5 and 6 in Lane and Ross 1994 and the descriptions on pp. 450-451).

and thus to increase the regime's approval (Gorbachev, 1987). However, these reforms instead were the trigger of a revolution as they unleashed social forces that brought about the dissolution of the USSR (Lane and Ross, 1994; Brown, 1997). These consequences were unintended by the leadership (see Gorbachev, 1987 p.17),[24] came to the surprise of most experts (as documented by Kuran 1991 and Lipset and Bence 1994) and are indeed counter to the predictions of the standard models of revolutions. How can popular reforms, such as Perestroika, trigger a revolution? Our model provides a possible answer to this unresolved question. In the case described in this section, the implementation of popular policies leads to increased dissent by moderates as it becomes easier for them to speak their minds when their views are closer to the new policy. In the USSR example, Yeltsin found it easier following Perestroika to express his critique since this critique was not considered as extreme by the regime after it had changed its own policy. This initial moderate critique then paved the way for more extreme dissidents and for extremer critique. 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 assume that individual dissent is triggered only by a great misalignment of preferences with the regime. Popular policies reduce such misalignment and hence cannot trigger a revolution in these models. If instead one looks at a model that differentiates between different levels of dissent, like ours, it follows that also moderates may start a revolution and hence that popular policies *can* be the trigger. In fact, possibly realizing that his reforms had triggered the dissent, Gorbachev tried at the end of his rule to undo them and instead strengthen the post of the executive presidency (Lane and Ross, 1994 p. 448). But this was evidently too late. Note that, in parallel to Perestroika, Gorbachev also implemented Glasnost (increased openness and freedom of speech).[25] In our model this is equivalent to a decrease in $\bar{K}$, which fosters more dissent. Thus, both these reforms had the effect of undermining his regime.

5.1.2 The Arab Spring in Egypt in 2011

Also in Egypt dissent was initiated by moderates. Prior to 2011, extreme opposition to Mubarak was harshly sanctioned hence to a large extent excluded from the public

[24]Gorbachev implemented Perestroika after a period of growing dissent. He was hoping that these reforms would put that dissent to rest (Gorbachev, 1996, p.349). Instead they did the opposite.

[25]A number of other policies were implemented at the same time as well, e.g., restricting alcohol which could be interpreted as an unpopular policy.

sphere.[26] Meanwhile, moderate opposition movements like Kefaya, which was founded by Egyptian intellectuals in 2004 to protest against Mubarak's intention to transfer power directly to his son Gamal, were allowed to run for parliament and protest on the streets.[27] Likewise, when the Arab Spring revolution broke 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, Gama'a al-Islamiyya and the Salafi movement) were not present in the protests initially. They only joined later, after Mubarak had been weakened by the initial protests and it was safer for them to express their views.[28] Once they joined, they were advocating the implementation of Sharia law thus challenging the claims of the initial protesters.

Just like in the USSR case, the Arab-Spring revolution in Egypt was two-sided essentially from the beginning. The protesters on the Tahrir Square consisted of some who suggested that Mubarak's regime was not sufficiently liberal and of others who said he was not sufficiently conservative and religious. Furthermore, while 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 of Figure 5) may have been conducive to the burst of the revolution, the later protests and elections revealed a different picture of the true preferences of Egyptian society (POMEPS, 2011).[29] It was revealed 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 5, where the average opinion after the shift is to 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.

In Appendix B we discuss some alternative and complementary explanations that

[26]Muslim-Brotherhood members were banned from running to parliament and many of them were arrested during the 2005 parliamentary elections campaign. However, some did run under other names.

[27]One of Kefaya's founding members, Hany Anan, even declared: "We are showing Egyptians that we can challenge the ruler, we can tell him we don't want you, that's enough, you go, and we can do this in public and still go back to our homes, maybe with some wounds or some bruises, but we still go home" (Saleh 2005). Other organizations were, however, banned.

[28]For example, 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).

[29]The shift in the population's preferences towards more liberalism/secularism may have been the result of the economy-wide growth (Davies 1962; Gurr 1970; Inglehart and Welzel 2006) or the increase in the level of education (see Goldstone 2011 and Campante and Chor 2012). The further potential effects of the education reforms are discussed in Appendix B.

are orthogonal to ours, such as those attributing the cause of the revolution to economic incentives.

6 All groups start the revolution

An interesting aspect that is absent from the analysis in most previous models is the characterization of how revolutions that are led by extremists (in the sense of who dissents the most) will evolve over time. This is exactly what distinguishes the third type of revolutions analyzed in this section (where $\alpha > 1$ and $\beta \geq 1$) from the type of revolution described in Section 4, which both share the leading role of extremists in the revolution. However, as opposed to the revolution type of Section 4, the type of revolution analyzed here is characterized by a moderate-to-extreme progress of public statements during the revolution. The full properties of this revolution are outlined 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 the implementation of a new policy only if it is unpopular.*
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.*
4. ***Dynamics of statements**: The most extreme types are the ones dissenting the most at all time periods and everyone increases their dissent over time.*

Proof. See Appendix A.5. ■

An important feature of this case is that $\beta > 1$ (S is convex), which represents a regime that is tolerant towards small dissent while punishing harshly larger dissent. This intuitively implies that each type will compromise between fully obeying the regime and speaking his mind: since 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. But the convexity of D also implies they are sensitive to large deviations from their bliss points, implying that extremists, whose bliss points are far from the regime, dissent more than moderates.[30] Hence, those leading the revolution will inevitably be the extremists. However, as dissenting extremely is sanctioned harshly, they will start off by expressing mild critique. When they do so, the approval and thus strength of the regime fall and these extremists then find it possible to express harsher critique. Meanwhile more moderate individuals increase their dissent too. This implies further weakening of the regime, enabling extremer dissent and so on. This way, the most extreme types lead the way during the revolution and continuously push the freedom of speech further, backed up from behind by moderates (points 3 and 4 of the proposition). This is an important difference between this class of revolutions and the revolutions described in Section 5. While both are characterized by statements becoming extremer over time, they differ in who the initiators are – moderates in Section 5 versus both groups here.

The fact that extremists dissent more than moderates do, further implies that it is great misalignment between an individual's ideology and the regime's policy that triggers dissent (point 1). Hence, implementation of unpopular policies, whereby the regime becomes more misaligned with the population, can trigger a revolution (point 2). This is depicted in Figure 7 where, in the top schedule, we are in a steady state with a somewhat left-biased regime. The regime then implements more left-biased, thus unpopular, policies (second schedule). This triggers more dissent, thereby weakening the regime, inducing more dissent and so on until the regime collapses in the bottom schedule.

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 (see phase diagram in Figure 6), unless the regime reacts on time by either increasing its force (e.g.,

[30]The best response of each type is unique for a given approval. The phase diagram of approval is depicted in Figure 6 and explained in Appendix Section A.5.

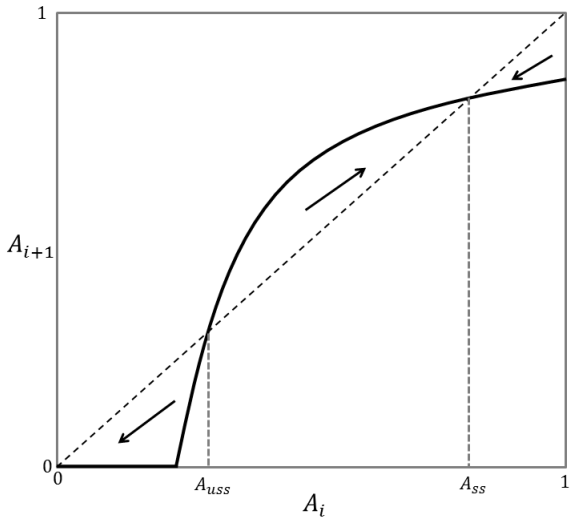

Figure 6: Stylized phase diagram for the case $\alpha > 1$ and $\beta \geq 1$ when the regime is biased. The solid line depicts the intertemporal-dynamics function $A_{i+1} = f(A_i)$ and the dashed line depicts 45-degree line where $A_{i+1} = A_i$. The vertical lines depict the stable (A_{ss}) and unstable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

by recruiting more troops) or implementing popular policies to appease the population.

7 Endogenous regime sanctioning

In this section we endogenize the sanctioning structure implemented by the regime. In particular, we aim to find which curvature of sanctioning a regime would choose if it could do so optimally. We give the regime extensive flexibility, assuming that it wishes to maximize its approval in any single period and can do so by choosing any $\beta_i > 0$. This means that β can change over time. The only constraint is that the regime takes its current strength K_{i} (or, put differently, the approval in the previous period A_{i-1}) as given. Formally, the regime solves the following problem:

$$\underset{\beta_i>0}{max} A_i$$

$$s.t.\ A_i = max\{0, 1 - \int_{t\in T} |x_i^*(t) - R|\, dt\},$$

$$x_i^*(t) = \arg\min_{x_i}\{L(x_i; t, K_i)\}$$

$$\text{and } K_i = \bar{K} A_{i-1}.$$

This is a mathematically difficult problem to solve for general distributions. Hence, to capture the main tension between what extremists and moderates do, we will analyze

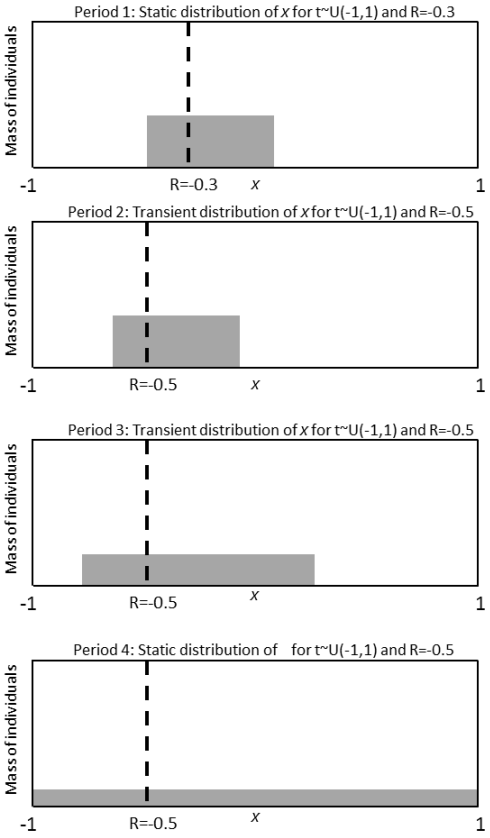

Figure 7: Distribution of stances over time in a stylized revolution starting with extremists dissenting moderately ($\alpha > 1$, $\beta \geq 1$). The regime starts at $R = -0.3$ and in the second period moves to $R = -0.5$ (which triggers the revolution) and stays there. The distribution of types is constant. For expositional purposes, the diagram depicts a case of $\alpha = \beta$.

a case where the regime policy $R = 0$ and half of the population are moderates, with $t = 1/2$, while the other half are extremists with $t = 1$.

Note that the regime's problem is equivalent to minimizing total dissent:

$$\min_{\beta_i>0} \int_{t\in T} |x_i^*(t)|\, dt. \tag{8}$$

We first show that the main result of the paper – the existence of three classes of revolutions – is robust to letting the regime choose its sanctioning endogenously. Before doing so, we need to slightly redefine what a revolution is. This is since, previously, dissent of all individuals was non-decreasing throughout a revolution. Now, with an endogenous β, dissent of one type can increase while it decreases for another type (due to a change in the regime's sanctioning between periods). The following definition takes this into account.

Definition 3. A revolution is a sequence of time periods in which aggregate dissent in the population (8) increases hence aggregate approval decreases.

In order to study what happens in a revolution we first need to know the properties of the state where the regime is stable.

Lemma 1. *Let $\bar{K} > 1$. Then, for any $\alpha > 0$, there exists a stable steady state where the regime's optimal sanctioning $\beta_i^* < 1$ and $x_i^*(t) = 0 \forall t$.*

Proof. See Appendix A.6.1. ■

The lemma establishes that a sufficiently forceful regime would choose $\beta^* < 1$, which results in zero dissent that guarantees the regime's stability. The following proposition outlines the endogenous sanctioning choice vis-a-vis our typology of revolutions.

Proposition 5. *Suppose the regime chooses in each period its sanctioning β_i^* to minimize dissent. Let $R = 0$, and suppose half the population has $t = 1/2$ and half has $t = 1$. Then, starting from $\bar{K} > 1$:*

1. *(**Extremists start the revolution**) If $\alpha \in (\underline{\alpha} \approx 0.53, 1)$, then there exists a shock to $\bar{K}$ such that a revolution starts with regime response $\beta^* < \alpha$, extremists dissenting and moderates staying silent;*

2. *(**Moderates start the revolution**) For any $\alpha < 1$, there exists a shock to $\bar{K}$ such that a revolution starts with regime response $\beta^* > \alpha$ and, if $\alpha > \tilde{\alpha} \approx 0.215$, extremists dissenting strictly less than moderates;*

3. *(**All groups start the revolution**) For any $\alpha > 1$, there exists a shock to $\bar{K}$ such that a revolution starts with regime response $\beta^* > 1$, where all groups start dissenting but the extremists dissent the most.*

Proof. See Appendix A.6. ■

Proposition 5 shows that, depending on the size of the shock to the regime's force, each of the three types of revolutions might be started. Most importantly, as indicated in point 2, a revolution started by moderates (the second class of revolutions) is a possible outcome also when the regime reacts optimally, provided that $\alpha < 1$. The restrictions in parts 1 and 2 of the proposition are partly due to the assumption of only two types of individuals (instead of a continuum) and in part due to technical difficulties in the analysis. According to simulations the results hold for all α. But ultimately, since we cannot prove it, it remains a conjecture. We now highlight the further properties of optimal sanctioning in a few illustrations.

7.1 Illustration: The dynamics of sanctioning when $\alpha < 1$

Here we illustrate how the regime's sanctioning changes dynamically depending on its current strength K_i (henceforth simply K) when $\alpha < 1$. As in Proposition 5, there are two equally-sized factions, moderates at $t = 1/2$ and extremists at $t = 1$, and a regime with $R = 0$. Figure 8 illustrates the collection of lemmas (Lemmas 10, 12 and 13 in the appendix) that jointly establish parts 1 and 2 of Proposition 5.

The upper panel in Figure 8 shows the regime's optimal choice of sanctioning β^* as a function of its current strength K. The middle panel shows how much each of the groups dissents for each K and its corresponding β^*, with larger circles representing extremists ($t = 1$) and smaller circles representing moderates ($t = 1/2$). The bottom panel shows the overall approval, which is a function of the dissent of both types.

It is most informative to study the figure from right to left, reflecting the dynamics of a revolution. Furthest to the right is the steady-state region, where $K > 1$. There the regime can induce complete silence by choosing a small β^*. For K somewhat smaller than 1, the regime can no longer induce complete silence by the extremists, only by the moderates. If the regime chooses $\beta < 1$, it induces silence by moderates but alienates the extremists. If it chooses $\beta > 1$ instead, the regime can induce compromise by the extremists, but the convexity implies weak incentives for moderates to not dissent.

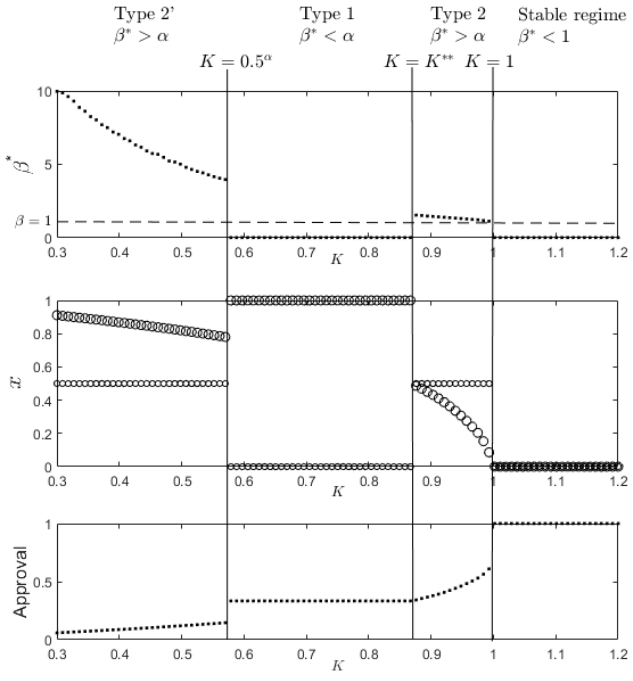

Figure 8: Illustration of endogenous β when $\alpha = 0.9 < 1$. Upper panel: β^* as function of K. Middle panel: the resulting dissent of each group. Lower panel: the resulting approval. $K^{**} \equiv \alpha ln(2) 2^{1/ln(2)-\alpha}$, see Lemma 8.

Hence, when K is somewhat smaller than 1, there is a trade off between incentivizing silence by moderates and incentivizing compromise by extremists. Which of these two considerations dominates boils down to which of them lowers dissent the most. Silencing moderates lowers dissent by 1/2. Hence focusing on compromise by extremists ($\beta > 1$) is worthwhile if and only if it reduces the extremists' dissent by more than 1/2. In the region $K^{**} < K < 1$, the compromising effect of extremists is strong enough and, hence, it is optimal to direct sanctioning at extremists by choosing $\beta^* > 1$.[31] In this region of K we thus have that extremists dissent *less* than moderates. Hence, should a revolution start following a small shock to the regime's strength from above 1 to just below it, this revolution will be started by moderates dissenting the most – a revolution of type 2.

Moving left in the figure, to the region where $(1/2)^{\alpha} < K < K^{**}$, the optimal sanction changes properties. The same trade off as before still applies – that between inducing silence by moderates ($\beta < 1$) or compromise by extremists ($\beta > 1$). But now,

[31] Note that we have described here the trade off under the assumption that a given size of dissent is equally harmful no matter which group does it. Should the groups be of different importance, or their sizes be unequal, or if the marginal "damage" of large dissent would be different than that of small dissent, then the value of K^{**} would change accordingly, but the logic would not.

when $K < K^{**}$, it is no longer possible to induce sufficient compromise by extremists (to lower their dissent by more than 1/2), while it is still possible to induce full silence by moderates since $(1/2)^{\alpha} < K$. Hence, the regime chooses $\beta^* < 1$, focusing fully on moderates and now alienating the extremists. Should a revolution start following a large shock to the regime's strength from above 1 to $(1/2)^{\alpha} < K < K^{**}$, that revolution will be started by extremists dissenting extremely while the moderates will remain silent – a revolution of type 1.

Finally, in the region where $K < (1/2)^{\alpha}$, the regime can no longer silence the moderates. Choosing $\beta < 1$ then becomes meaningless, as it has no effect on anyone. What is left for the regime to do is to induce some compromise by the extremists by choosing $\beta^* > 1$, leaving the moderates to speak their minds. In this region β^* is increasing as K becomes smaller. This is since when K becomes smaller, the regime has to ramp up the *marginal* sanctioning by choosing a higher β. In this region of K a revolution follows the pattern of type 2 in terms of sanctioning and of type 3 in terms of dissent (hence we denote it in the figure by type 2'): Extremists are dissenting the most and are gradually increasing their dissent. We conjecture that, had we considered a continuous distribution of types, the dissent would have followed a type-2 revolution.

The above analysis is based on the regime's strength K, which in itself is a function of its approval A and force $\bar{K}$. Since $\bar{K}$ is exogenous, it is in principle possible to find $\bar{K}$ such that any K in Figure 8 regenerates itself, i.e., a steady state. Thus, there may exist multiple steady states.

This has a few implications. First, a regime can be stable while there exists dissent in society, and this dissent may be led by moderates (if $K > K^{**}$ is stable) or by extremists (if $K < K^{**}$ is stable). Another implication is that a type-2 revolution may start following a small shock to K (so that it drops to a value a bit below 1), but then it may transform into a type-1 revolution. That is, the extremists may take over the revolution while moderates go back to silence. This also means that dissent by a particular group may be non-monotonic during a revolution up until the final stage (when $K < (1/2)^{\alpha}$). At this final stage the dissent only increases, as can be seen in the left region of the middle panel of Figure 8. The figure also shows that not only dissent, but also the regime's sanctioning structure may shift non-monotonically throughout the revolution. In particular, the regime may focus initially on extremists, then focus on moderates, then again focus on extremists.

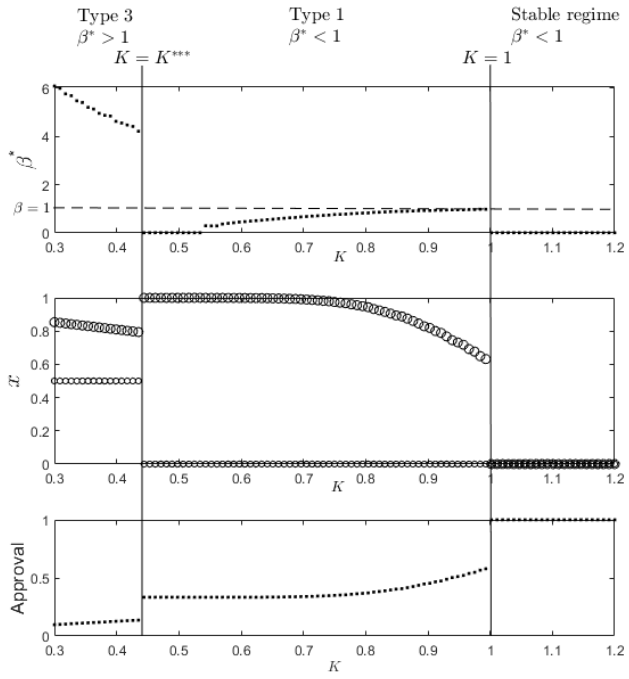

Figure 9: Endogenous β when $\alpha = 1.2 > 1$. Upper panel: β^* as function of K. Middle panel: the resulting dissent of each group. Lower panel: the resulting approval. K^{***} is defined in Lemma 14.

7.2 Illustration: The dynamics of sanctioning when $\alpha > 1$

Here we illustrate how the regime's sanctioning changes dynamically depending on its current strength K when $\alpha > 1$. Figure 9 illustrates the results of Lemma 14 in the appendix and parts 1 and 3 of Proposition 5.

Recall that, when $\alpha > 1$, all types compromise a little bit, as this is costless in terms of D, but larger compromise is increasingly costly. This means that extremists will always dissent (weakly) more than moderates and that inducing complete silence by extremists requires a heavy punishment – large K. But when K *is* large, the regime can achieve complete silence by all by choosing $\beta^* < 1$. This is the rightmost region in the figure, where $K > 1$. When K is smaller than 1, the regime can no longer induce complete silence by extremists and has a trade off. If it chooses $\beta > 1$, it provides a strong incentive for extremists to compromise, but a weak incentive for moderates to compromise. If, on the other hand, it chooses $\beta < 1$, it provides a strong incentive for moderates to stay completely silent, but a weaker incentive for extremists to compromise a little. When K is close to 1, the latter is better for the regime ($\beta^* < 1$), as the regime is strong enough to induce compromise by the extremists even when $\beta < 1$. However,

there is considerable dissent by extremists, and the smaller is K the stronger is this dissent. This can be seen in the middle region of the middle panel of Figure 9 where, going to the left, the larger circles representing the extremists move towards 1.

For a sufficiently small K (smaller than roughly 0.45, the leftmost region in Figure 9), it is no longer possible to induce silence by the moderates by choosing $\beta^* < 1$. Given that also the compromise of the extremists is minimal, the regime changes focus. By choosing $\beta^* > 1$, it ignores the moderates and instead ensures at least some compromise by extremists. In this region β^* increases as K falls, which captures that a weaker regime has to resort to extreme marginal punishment to get any compromise by the extremists.

Overall it can be noted that β^* is non-monotonic as K falls. β^* is falling from 1 to 0 in the middle region as K decreases, but then jumps to around four and from there increases to infinity in the leftmost region. The above logic regarding what the regime is focusing on also implies that the dissent of the extremists is non-monotonic. It first gradually increases as K decreases, but then discontinuously drops when moving from the middle to the left region. It is at that drop that the regime refocuses on getting compromise by the extremists.

The comments regarding the possibility of multiple steady states (see previous subsection) apply also here.

7.3 Illustration: Population shares and optimal sanctioning

Up till here, we assumed that the moderates ($t = 1/2$) and the extremists ($t = 1$) are of equal size. This was done for tractability of the formal analysis. In this short subsection we wish to study the effect of the relative sizes of the two groups. For this purpose, suppose that a share μ of the population are moderates with $t = 1/2$ and a share $1 - \mu$ are extremists. To highlight the effect of μ we restrict the analysis to $\alpha = 1$.

Proposition 6. *Let $R = 0$, $\alpha = 1$, and suppose that all individuals have either $t = 1/2$ or $t = 1$. Then, for any K, β^* weakly decreases in the share of moderates μ.*

Proof. See Appendix A.6.5. ■

Proposition 6 highlights an intuitive result: the greater is the share of moderates in society (or the more popular the regime's ideology is), the better it is for the regime to

concentrate the sanctioning on small dissent rather than on larger dissent, i.e., to prefer a concave sanctioning over a convex one.

8 Concluding discussion

This paper is the first to explain (1) why revolutions sometimes start with moderate opponents of the regime and (2) why the implementation of popular policies sometimes triggers a revolution. It answers these questions by relaxing a seemingly innocuous assumption of most standard models of revolutions and mass protests (starting from Granovetter 1978 and Kuran 1989). Namely, that potential participants face a binary choice: either obey the regime or fully participate in a revolution, i.e., they cannot choose any middle ground. Relaxing this assumption and allowing for reasonably general functional forms turns out to fundamentally alter the predictions about the dynamics of revolutions. The more general framework presented in the current paper yields a typology with three types of popular revolutions and mass protests: 1) revolutions started by extremists; 2) revolutions started by moderates; and 3) revolutions starting with all groups increasing their dissent. Earlier models invariably predict that any revolution will be initiated by ideological extremists and are silent about the extent to which each individual will dissent and how this will change over time. Hence, these models cannot account for the second type of revolutions, whose dynamics fit that of the Arab spring in Egypt in 2011 and the fall of Communism in 1989-91.

The overarching pattern is that which *faction* (extremists or moderates) will start the revolution is determined by the curvature of the cost of deviating from the individual bliss point. A convex ideological cost makes it costless for the individual to deviate a little from the bliss point hence moderates will stay silent while extremists are more prone to start a revolution since for them it is too costly to stay completely silent. On the other hand, a concave ideological cost induces people to either speak their minds or stay completely silent. Hence moderates, who are less heavily punished when speaking their minds, are more inclined to protest, while extremists (who are more heavily punished when speaking their minds) are silent. This implies that moderates may initiate a revolution while extremists will only join later in time, when the regime's strength has subsided. This thus provides an answer to our first research question. Furthermore, this logic implies that great misalignment with the regime's policy silences an individual. Hence, if the regime's policy is misaligned with large parts of society there will be less protest. A popular policy may then trigger a revolution because, indirectly, it spurs

more people to become "moderate" hence speak their minds. This provides an answer to our second research question.

Previous research and anecdotal evidence suggest that a variety of sanctioning regimes may exist in practice (Milani 1988 p.122, p.197; Saleh, 2005; Hermann et al 2008; Michaeli and Spiro, 2015) and that societies and ideological issues are heterogeneous in terms of the ideological cost associated with them (Kendall et al., 2015; Chen et al, 2020; Abeler et al. 2019). In particular, there is evidence suggesting that concave ideological costs are prevalent in political and moral settings (Kendall et al., 2015; Kajackaite and Gneezy 2017; Gneezy et al. 2018; Krupka and Weber, 2013). However, to our knowledge, there exists no systematic comparison of the curvature of ideological and sanctioning costs across countries, and in particular no available data applicable to the examples of Egypt and the USSR that we have provided. Nevertheless, an emerging body of research has started developing ways of measuring the curvature of bliss-point deviations (see e.g. Baranski et al. 2020). If such methods become feasible to conduct across societies and possibly across time, then our theory can be used not only to rationalize the puzzling events presented but can also be used for testing Corollary 1.

To transparently focus on the mechanism that drives our results, we have deliberately kept the model as simple and as close as possible to the canonical model by Kuran (1989). There are of course numerous additions one can make to our model. We study one such extension in the paper: allowing the regime to change its sanctioning structure to maximize its approval. This extension provides insights into how a rational and fully flexible regime could adapt its sanctioning over time during a revolution. Most importantly, the mechanism underlying the answer to our research questions remains – when individuals have concave ideological costs, the regime's best response to a shock may still be to choose sanctioning that induces moderates to start the revolution. There are clearly other extensions to consider, including intervention by outside forces; conflicts between different revolutionary groups about the targets of the revolution; and differences between individuals other than ideological. We have no reason to believe such additions change our conclusions. But ultimately we view it as a subject for future research.

References

[1] Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. Econometrica, 87(4), 1115-1153.

[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.

[3] Aminzade, R., (2001). *Silence and voice in the study of contentious politics*. Cambridge U. Press.

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

[5] Armbrust, W., (2011). "Egypt: A Revolution against Neoliberalism?". AlJazeera 24th Feburary.

[6] Bala, V., & Goyal, S. (2001). Conformism and diversity under social learning. *Econ. theory*, 17(1), 101-120.

[7] Baranski, A., Haas, N., & Morton, R. (2020). "Majoritarian Bargaining over Budgetary Divisions and Policy". NYU Abu Dhabi, Working Paper, 52.

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

[9] Bernheim, D.B., (1994), "A Theory of Conformity", *J. of Political Economy,* 102(5), 841-877.

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

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

[12] Brudny, Y.M., (1993)."The Dynamics of 'Democratic Russia', 1990-1993". *Post-Soviet Affairs*, 9(2), 141-170.

[13] Bueno De Mesquita, E. (2010). "Regime change and revolutionary entrepreneurs". *Amer. Pol. Sci. Rev.*, 104(03), 446-466.

[14] Bursztyn, L., Egorov, G., Enikolopov, R. and Petrova, M., (2019). "Social media and xenophobia: evidence from Russia". WP No. w26567, National Bureau of Economic Research.

[15] Campante, F.R. and Chor, D., (2012). "Why was the Arab world poised for revolution? Schooling, economic opportunities, and the Arab Spring". *The J. of Econ. Perspectives*, 26(2), 167-187.

[16] Chen, D. L., Michaeli, M., & Spiro, D. (2019). "Non-confrontational extremists", WP.

[17] Chen, H. and Suen, W., (2016). "Falling dominoes: A theory of rare events and crisis contagion". *Amer. Econ. J.: Microeconomics*, 8(1), 228-255.

[18] Chen, H. and Suen, W., (2017). "Aspiring for Change: A Theory of Middle Class Activism". *The Econ. J.*. Vol 127, Iss.603, 1318-1347.

[19] Chen, H. and Suen, W., (2020, forthcoming). "Radicalism in Mass Movements: Asymmetric Information and Agenda Escalation." *American Political Science Review.*

[20] Chenoweth, E., Pinckney, J. and Lewis, O.A. (2017). "Nonviolent and Violent Campaigns and Outcomes Dataset, v. 3.0". University of Denver.

[21] Chwe, M. S. Y. (1999). Structure and strategy in collective action. *Amer. J. of Soc.*, 105(1), 128-156.

[22] Corsetti, G., Dasgupta, A., Morris, S. and Shin, H.S., (2004). "Does one Soros make a difference? A theory of currency crises with large and small traders". *The Rev. of Econ. Stud.*, 71(1), pp.87-113.

[23] Dagaev, D., Lamberova, N., & Sobolev, A. (2019). Stability of revolutionary governments in the face of mass protest. *European Journal of Political Economy*, 60, 101812.

[24] Dahlum, S. and Knutsen, C.H., (2017). "Democracy by demand? Reinvestigating the effect of self-expression values on political regime type". *British J. of Pol. Sci.*, 47(2), 437-461.

[25] Davies, J.C. (1962). "Towards a Theory of Revolution." *Amer. Soc. Rev.*. 27(1):5–19.

[26] Desai, R.M, Olofsgård, A and Yousef, T.M. (2019). "Nonviolence and Violence in Anti-Regime Politics A Signaling Model with Evidence from the Arab World", mimeo Georgetown University.

[27] Edmond, C. (2013), "Information Manipulation, Coordination, and Regime Change", *Rev. of Econ. Stud.*, Vol. 80, 1422–1458.

[28] Edwards, B. and Gillham, P.F., 2013. "Resource mobilization theory" in *The Wiley-Blackwell Encyclopedia of Social and Political Movements*.

[29] ElTantawy, N. and Wiest, J.B., (2011). "The Arab spring: Social media in the Egyptian revolution: reconsidering resource mobilization theory". *International J. of Communication*, 5, 18.

[30] Enikolopov, R., Makarin, A., Petrova, M. and Polishchuk, L., 2017. "Social image, networks, and protest participation". WP March 24, 2017).

[31] Esteban, J.; Ray, D. (2001). "Collective action and the group size paradox." *Amer. Pol. Sci. Assoc.* Vol. 95, No. 03, 663-672.

[32] Fischbacher, U. and Föllmi-Heusi, F., 2013. "Lies in disguise—an experimental study on cheating". *Journal of the European Economic Association*, 11(3), pp.525-547.

[33] Gates, S., Hegre, H., Jones, M.P. and Strand, H. (2006). "Institutional Inconsistency and Political Instability: Polity Duration, 1800–2000." *Amer. J. of Pol. Sci.* 50(4): 893–908.

[34] Gibson, J.L. (1997). "Mass Opposition to the Soviet Putsch of August 1991: Collective Action, Rational Choice, and Democratic Values in the Former Soviet Union". *The Amer. Pol. Sci. Rev.*, Vol 91, Iss 3 (Sep., 1997), 671-684.

[35] Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. American Economic Review, 108(2), 419-53.

[36] Goldstone, J. A. (2001). "Toward a fourth generation of revolutionary theory". *Ann. Rev. of Pol. Sci.*, 4, 139-187.

[37] Goldstone, J.A., (2011). "Understanding the revolutions of 2011: weakness and resilience in Middle Eastern autocracies". Foreign Affairs., 90, 8.

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

[39] Gorbachev, M. (1996), *"Mikhail Gorbachev: Memoirs"*. Doubleday.

[40] Granovetter, M. (1978), "Threshold Models of Collective Behavior", *The Amer. J. of Soc.,* 83(6): 1420-1443.

[41] Gurr, T.R. (1970). *"Why Men Rebel"*. Princeton, NJ: Princeton University Press.

[42] Gurr, T.R. (1974). "Persistence and Change in Political Systems, 1800–1971." *Amer. Pol. Sci. Rev.* 68(4): 1482–1504.

[43] Hegre, H. and Sambanis, N. (2006). "Sensitivity analysis of empirical results on civil war onset". *J. of conflict resolution*, 50(4), 508-535.

[44] Herrmann, B., Thöni, C., Gächter, S., (2008). Antisocial punishment across societies. Science 319 (5868), 1362–1367.

[45] Inglehart, R. and Welzel, C. (2005). *"Modernization, Cultural Change and Democracy—The Human Development Sequence"*. Cambridge: Cambridge University Press.

[46] Jenkins, J.C., (1983). "Resource mobilization theory and the study of social movements". *Ann. Rev. of Soc.*, 9(1), 527-553.

[47] Jia, R., (2014). "Weather shocks, sweet potatoes and peasant revolts in historical China". *The Econ. J.*, 124(575), 92-118.

[48] Kajackaite, A. and U. Gneezy (2017). Incentives and cheating. Games and Economic Behavior 102, 433–444.

[49] Kandil, H., (2012). "Why did the Egyptian middle class march to Tahrir Square?". *Mediterranean Politics*, 17(2), 197-215.

[50] Kendall, C., T. Nannicini, and F. Trebbi (2015, January). How Do Voters Respond to Information? Evidence from a Randomized Campaign. American Economic Review 105(1), 322–53.

[51] Knutsen, C.H., (2014). "Income growth and revolutions". *Social Science Quarterly*, 95(4), 920-937.

[52] Knutsen, C.H. and Nygård, H.M., (2015). "Institutional Characteristics and Regime Survival: Why Are Semi-Democracies Less Durable Than Autocracies and Democracies?". *Amer. J. of Pol. Sci.*, 59(3), 656-670.

[53] Korotayev, A. and Zinkina, J.V., (2011). "Egyptian revolution: A demographic structural analysis". *Entelequia. Revista Interdisciplinar*, 13(2011), 139-169.

[54] Krupka, E.L. and Weber, R.A., (2013). "Identifying social norms using coordination games: Why does dictator game sharing vary?". *Journal of the European Economic Association*, 11(3), pp.495-524.

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

[56] Kuran, T., (1991), "Now out of Never, The element of surprise in the east European revolution of 1989", *World Politics*, 44(1), 7-48.

[57] Kuran, T., (1995), "The Inevitability of Future Revolutionary Surprises," *The Amer. J. of Soc.*, 100(6), 1528-1551.

[58] Kuran, T., & Sandholm, W. H. (2008). "Cultural integration and its discontents". *The Rev. of Econ. Stud.*, 75(1), 201-228.

[59] Lane, D. and Ross, C. (1994). "The Social Background and Political Allegiance of the Political Elite of the Supreme Soviet of the USSR: The Terminal Stage, 1984 to 1991". *Europe-Asia Stud.*, 46(3), 437-463

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

[61] Lim, M., (2012). "Clicks, cabs, and coffee houses: Social media and oppositional movements in Egypt, 2004–2011". *J. of communication*, 62(2), 231-248.

[62] Lipset, S.M. and Bence, G., 1994. "Anticipations of the Failure of Communism". *Theory and Society*, 23(2), 169-210.

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

[64] Manski, C.F.,Mayshar, J. (2003)"Private Incentives and Social Interactions: Fertility Puzzles in Israel," *J. of the Eur. Econ. Assoc.*, 1(1), 181-211.

[65] McAdam, D., (1986). "Recruitment to high-risk activism: The case of freedom summer". *Amer. J. of Soc.*, 92(1), 64-90.

[66] Meyer, D.S., (1993). Peace protest and policy: explaining the rise and decline of antinuclear movements in postwar America. *Policy Studies Journal*, 21(1), pp.35-51.

[67] Meyer, D.S., (2004). Protest and political opportunities. *Annu. Rev. Sociol.*, 30, pp.125-145.

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

[69] Michaeli, M. and Spiro, D. (2017). "From Peer Pressure to Biased Norms" *Amer. Econ. J.: Microeconomics*, 9(1): 152-216.

[70] Milani, M.M., (1988, 1994). *The making of Iran's Islamic revolution: from monarchy to Islamic republic*. Westview Pr.

[71] Moaddel, M. (1992). "Ideology as episodic discourse: the case of the Iranian revolution". *Amer. Soc. Rev.*, 353-379.

[72] Narciso, G. and Severgnini, B., (2016). "The Deep Roots of Rebellion: Evidence from the Irish Revolution" Working Paper No. tep2216. Trinity College Dublin, Department of Economics.

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

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

[75] Orwell, G., (1949). *1984*. Everyman's Library.

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

[77] Passarelli, F., & Tabellini, G. (2017). Emotions and political unrest. *Journal of Political Economy*, 125(3), 903-946.

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

[79] POMEPS (2011) "Arab uprisings: the state of the Egyptian revolution", Project on Middle East Pol. Sci. Briefing no. 6.

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

[81] Rubin, J. (2014). "Centralized institutions and cascades". *J. of Comparative Econ.* 42(2), 340–357

[82] Saleh, H. (2005), "Analysis: Re-Birth of Egyptian Politics", BBC News, 5 September 2005. http://news.bbc.co.uk/2/hi/middle_east/4216750.stm. Accessed May 2017.

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

[84] Shadmehr, M. (2015). "Extremism in Revolutionary Movements". *Games and Econ. Behavior*, 94, .97-121.

[85] Shadmehr, M., & Bernhardt, D. (2011). Collective action with uncertain payoffs: coordination, public signals, and punishment dilemmas. *American Political Science Review*, 829-851.

[86] van Stekelenburg, J. and Klandermans, B., (2017). "Individuals in movements: A social psychology of contention". In *Handbook of social movements across disciplines* (pp. 103-139). Springer, Cham.

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

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

[89] Urban, M. and Gelman, V., (1997). "The development of political parties in Russia" in Democratic changes and authoritarian reactions in Russia, Ukraine, Belarus, and Moldova, pp.175-219.

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

[91] Young, H. P. (1993). "The evolution of conventions". *Econometrica*, 57-84.

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

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

ONLINE APPENDIX

A Analytical derivations and proofs

A.1 Individual stances

The individual minimizes the loss function given by (3), (1) and (2) when $K > 0$. Using the implicit function theorem we get the following derivatives of $x^*(t)$ in inner solutions:

$$\frac{dx^*}{dt} = \frac{D''(t - x^*)}{S''(x^*) + D''(t - x^*)} \tag{9}$$

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.1.1 Case (1): $\max\{\alpha, \beta\} \leq 1$

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 extremum point is a maximum. The corner solutions are then either $L(x = R) = |t - R|^\alpha$ or $L(x = t) = K|t - R|^\beta$.[32] When $\beta < \alpha$ this implies that $L(x = R) < L(x = t)$ iff $|t - R| < \Delta$, and so $x^*(t) = t$ iff $|t - R| \geq \Delta$, and $x^*(t) = R$ iff $|t - R| < \Delta$. When $\alpha < \beta$ the converse holds,[33] with $x^*(t) = t$ iff $|t - R| \leq \Delta$, and $x^*(t) = R$ iff $|t - R| > \Delta$.

A.1.2 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 $|x^*(t) - R|$ 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 $x^* = R$. In order to find the global minimum for a type t, we first need to investigate the behavior of $L(x, t)$ at $x = t$ and $x = R$.

$$L'(x, t) = -\alpha(t - x)^{\alpha-1} + \beta K(x - R)^{\beta-1} \tag{10}$$

[32] Strictly speaking, R is not a corner solution, but it is clear that the solutions to the optimization problem of the individual are in the range $[R, t]$.

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

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

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

Define

$$\Phi(x) \equiv (t-x)^{\alpha-1} (x-R)^{1-\beta} \ . \tag{12}$$

For the existence of an inner min point for a given t it is necessary that $\Phi(x) = \beta K/\alpha$ for some $x \in]R,t[$. Note that as $t \to R$ both $(t-x)^{\alpha-1}$ and $(x-R)^{1-\beta}$ approach zero implying $\Phi(x) < \beta K/\alpha$ for all $x \in]R,t[$. Hence types with sufficiently small $|t-R|$ do not have an inner local min point and they choose $x^* = R$. For sufficiently large $|t-R|$ it may be that $\Phi(x) = \beta K/\alpha$ for some $x \in]R,t[$ which we investigate next. Note that, for given t, $\Phi(x)$ is strictly positive in $]R,t[$, and that $\Phi(x,t) = 0$ at both edges of the range (i.e. at $x = R$ and at $x = t$). This means that $\Phi(x)$ has at least one local maximum in $]R,t[$. We now proceed to check whether this local maximum is unique:

$$\Phi'(x) = (t-x)^{\alpha-2} (x-R)^{-\beta} [(1-\beta)(t-x) - (\alpha-1)(x-R)] \tag{13}$$

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

$$L(R, t_h) - L(t_h, t_h) = |t_h - R|^\alpha - K |t_h - R|^\beta , \tag{14}$$

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 $|x^*(t) - R|$ is increasing in t. First note that $\Phi(x)$ is increasing in t, so if there exists a local minimum of $L(x, t_0)$ for some t_0, then there exists a local minimum of $L(x, t)$ for $t > t_0$ too. Also note that for all $x \in]R, t[\Phi(x)$ is increasing in t and that $\lim_{t\to\infty} \Phi(x, t) = \infty > \beta K/\alpha$, 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 $\Delta L \equiv L(R, t) - L(\tilde{x}, t) = (t - R)^\alpha - \left[(t - \tilde{x})^\alpha + K(\tilde{x} - R)^\beta\right]$, where $\tilde{x}$ is the stance at which $L(x, t)$ 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. Differentiating ΔL with respect to t and using the first-order condition (11) yields

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

Differentiating once more

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

By equation (9) we have that $\frac{d\tilde{x}}{dt} > 1$ in an inner solution when S 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 $\frac{dx^*}{dt} > 1$, and so $|x^*(t) - R|$ is increasing in t at the range of inner solutions.

A.1.3 Case (3): $\alpha < 1 < \beta$

The analysis here, performed for $t \geq R$, is very similar to that of Section A.1.2 above. We will first show that the only relevant corner solution is $x^* = t$, then that types close to the regime choose this corner solution. By (10) we have $L'(R,t) < 0$ and $L'(t,t) < 0$ since $\alpha < 1$. Therefore $x = t$ may be a solution to the minimization problem while $x = R$ is not. The candidate solution $x = t$ will now be compared to potential local minima in the range $[R,t]$. In inner solutions, (11) holds. Since $\alpha < 1$ and $\beta > 1$, it follows that $\Phi > \beta K/\alpha$ (see (12)) for all x when t is sufficiently small and K is finite. Hence, sufficiently small t do not have an inner local min point which implies $x^* = t$ is the global optimum for these types. Notice that $\Phi(x)$ is strictly positive in $]R,t[$, and that $\Phi(x) \to \infty$ at both edges of the range (i.e. at $x = R$ and at $x = t$). This means that $\Phi(x)$ has at least one local minimum in $]R,t[$. Analyzing (13) like in Section A.1.2 above, while noting that this time $[(1-\beta)(t-x) - (\alpha-1)(x-R)]$ is negative at $x = R$ and positive at $x = t$, we get a unique local minimum of $\Phi(x)$ in $]R,t[$ which, in case it is smaller than $\beta K/\alpha$, implies that $L(x,t)$ has exactly two extrema in the range $]R,t[$. From the negative values of $L'(x,t)$ at the edges of this range we finally conclude that the first extremum (where $\Phi(x)$ is falling) is a minimum point of $L(x,t)$, and the second extremum (where $\Phi(x)$ is rising) is a maximum point of $L(x,t)$. The global minimum of $L(x,t)$ is therefore either this local minimum (i.e. an inner solution), or $x = t$ (i.e. a corner solution). If however the value of $\Phi(x)$ at its local minimum point is larger than $\beta K/\alpha$, then there is no local extremum to $L(x,t)$ in the range $]R,t[$, and therefore $x = t$ is the solution to the minimization problem. Using the same logic and technique as in Section A.1.2, we get that $L(R,t_h) - L(t_h,t_h)$ (from (14)) is strictly negative when $|t_h - R| > \Delta = K^{\frac{1}{\alpha-\beta}}$ and $\alpha < \beta$, implying that t_h does not choose the corner solution of $t = t_h$, hence must choose an inner solution. Finally, it can be shown using the same steps of Section A.1.2 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; and that in the range of inner solutions $|x^*(t) - R|$ is decreasing in t.

A.1.4 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 $t \geq R$.

$$-\alpha (t - x)^{\alpha - 1} + \beta K (x - R)^{\beta - 1} = 0 \tag{15}$$

$$(\alpha - 1)\, \alpha (t - x)^{\alpha - 2} + (\beta - 1)\, \beta K (x - R)^{\beta - 2} > 0 \tag{16}$$

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 (15) and (16). Moreover, by applying the implicit function theorem to equation (15), we get that $dx^*/dt > 0$, hence $|x^*(t) - R|$ is strictly increasing in the distance to the regime. $1 = \beta < \alpha$: It is easy to verify that types sufficiently close to the regime choose $x^*(t) = R$ (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 $dx^*/dt = 0$. For the subrange with inner solutions using $\beta = 1$ and $\alpha > 1$ in equation (9) implies that $dx^*/dt = 1$ and hence $|x^*(t) - R|$ 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 $x^*(t) = t$, while types sufficiently far from the regime choose the same inner solution x s.t. $S'(|x - R|) = 1$ $(= D')$. It thus follows that $|x^*(t) - R|$ is first increasing in the distance from the regime and then it stays constant.

A.2 Proof of Proposition 1

A.2.1 Part 1

We start by showing that initially – i.e., in the steady state – the most dissenting types are extremists (i.e., $\max |x^*(t) - R|$ is achieved for $t = \arg\max_t |t - R|$). For $\alpha \leq 1$ this follows immediately from Section A.1.1. If instead $\alpha > 1$, we know from Section A.1.2 that if the range of types is not sufficiently broad, then $x^*(t) = R$ 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.1.2 further tells us that $x^*(t)$ is increasing in the subrange of types with inner solutions, implying that $\max |x^*(t) - R|$ 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.1.1 and A.1.2 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 ($x^*(t) = R$) but is indifferent between R and some $x \neq R$ ($x = t$ 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 S 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 S 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.2.2 Part 2

We start by showing that initially – i.e., in the steady state – the most dissenting types are moderates (i.e., $\max|x^*(t) - R|$ is achieved for $t \neq \arg\max_t |t - R|$). If $\beta \leq 1$, we know from Section A.1.1 that there exists a distance from the regime, $\Delta = K^{\frac{1}{\alpha-\beta}}$, such that a type at that distance chooses $x^*(t) = t$ 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 $x^*(t) = t$ for everyone hence the regime does not exist), this immediately implies that $\max|x^*(t) - R| = \Delta$ is achieved for $t = R \pm \Delta \neq \arg\max_t |t - R|$. Alternatively, if $\beta > 1$, we know from Section A.1.3 that if the range of types is not sufficiently broad, then $x^*(t) = t$ 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.1.3 further tells us that $x^*(t)$ is decreasing in the subrange of types with inner solutions, implying that $\max|x^*(t) - R|$ is achieved for $t \neq \arg\max_t |t - R|$. To see that, as the revolution evolves, more extreme types (compared to $\arg\max_t |x^*(t) - R|$ 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 $x^*(t) = t$) must have had a different solution at time i ($x_i^*(t) = R$ if $\beta \leq 1$, or an inner solution if $\beta > 1$), implying that t is further away from the regime (= a more extreme type) than the type who was most dissenting at time i (who is more extreme than the one most dissenting at time $i-1$ and so on until we reach the steady state).

A.2.3 Part 3

That initially – i.e., in the steady state – the most dissenting types are extremists (i.e., $\max |x^*(t) - R|$ is achieved for $t = \arg\max_t |t - R|$), follows immediately from Section A.1.4, where we show that $|x^*(t) - R|$ 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.1.4 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.3 Proof of Proposition 2

We first outline some helpful analytical results and then prove the actual proposition. First note from Section A.1.1 that $x^*(t)$ 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 $A_{i+1} = f(A_{i+1})$.

In what follows, the phase diagram in Figure 2 may be a useful aid.

Lemma 2. *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) If $R = 0$ then $f(A_i)$ is convex for $A_i > 0$. 4) If $R \neq 0$ 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 $\bar{K}_{c_1}$ such that a stable steady state with a regime and $A_{ss} > 0$ exists iff $\bar{K} > \bar{K}_{c_1}$. 9)$\bar{K}_{c_1}$ is increasing in $|R|$.*

Proof. From Section A.1.1 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 (6)

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

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

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

Hence we get

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

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

$$1 \leq R^2 - \Delta_i^2 + 1 \Leftrightarrow \Delta_i \leq -R.$$

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 $\Psi\left(x_i^*; R, A_i\right) \geq 1$ may hold also for some Δ_i in the middle region. Checking when this happens we get

$$\Delta_i \frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2 = 1 \Rightarrow \ldots \Rightarrow = \sqrt{\left(R - \left(1 + \sqrt{2}\right)\right)\left(R - \left(1 - \sqrt{2}\right)\right)},$$

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

get

$$A_{i+1} \equiv f\left(R, A_i\right) = \begin{cases} 0 & \text{when } 0 \leq \Delta_i \leq -R \\ 1 - \left(R^2 - \Delta_i^2 + 1\right) & \text{when } -R < \Delta_i \leq 1 + R \\ 1 - \left(\frac{1}{2} - R - \frac{1}{2}\Delta_i^2 + \frac{1}{2}R^2\right) & \text{when } 1 + R < \Delta_i < 1 - R \\ 1 & \text{when } 1 - R \leq \Delta_i \end{cases} . \quad (17)$$

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

$$\Delta_i = \left(\bar{K} A_i\right)^{\frac{1}{\alpha - \beta}}, \quad (18)$$

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 2 shows and will now be proved, the two middle regions are convex. Using (17) and (18)

$$\frac{df}{dA_i} = \left\{ \begin{array}{ll} \frac{2}{\alpha-\beta}\Delta_i^2 A_i^{-1} & \text{when } \Delta_i \leq 1 + R \\ \frac{1}{\alpha-\beta}\Delta_i^2 A_i^{-1} & \text{when } 1 + R < \Delta_i < 1 - R \end{array} \right\} > 0$$

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

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 $A_{i+1} = f\left(A_i\right)$ is flat at zero it means that the first intersection is unstable, the next is stable, next unstable and next stable. Using (17) and (18)

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

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

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

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: $A_{i+1}|_{\Delta_i=1+R} \geq A_i|_{\Delta_i=1+R} \Leftrightarrow \{\text{Using (17) and (18)}\} \Leftrightarrow 1+2R \geq (1+R)^{\alpha-\beta}/\bar{K}$. As the RHS is positive, this inequality can hold only if $R \in [-1/2, 0]$ 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 $A_i = 1$)$\Leftrightarrow \{\text{Using (18)}\} \Leftrightarrow 1-R \leq \bar{K}^{\frac{1}{\alpha-\beta}} \Leftrightarrow \bar{K} \geq (1-R)^{\alpha-\beta}$. Denote the smallest $\bar{K}$ fulfilling one of these conditions by $\bar{K}_{c_1}$. Thus follows part (8), and it can be verified that $\bar{K}_{c_1}$ increases in $|R|$ (proving part (9)). ■

Proof of Proposition 2 Part (1): follows from parts (8) and (9) of Lemma 2. Part (2): From definition 1 it follows that any convergence to a lower steady state constitutes a revolution because the approval falls over several periods. Hence 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 2), and in particular if the shock is such that $\bar{K}$ goes below $\bar{K}_{c_1}$, A converges to zero and the regime completely falls (part (8) of Lemma 2). Finally, implementation of unpopular policies means that $|R|$ increases, and as a result the approval of the regime decreases (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 (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 (17) – see Figure 2) and (ii) Δ_i starts above $1+R$ (where $x(t)$ 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 ($x(t) = t$).

A.4 Proof of Proposition 3

We first outline some helpful analytical results and then prove the actual proposition. First note from Sections A.1.1-A.1.4 that $x^*(t)$ 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 $A_{i+1} = f(A_{i+1})$.

In what follows, the phase diagram in Figure 4 may be a useful aid.

Lemma 3. *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) If $R = 0$ then $f(A_i)$ is concave for $A_i > \varepsilon$. 4) If $R \neq 0$ 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 $\bar{K}_{c_2}$ such that a stable steady state with a regime and $A_{ss} > 0$ exists iff $\bar{K} > \bar{K}_{c_2}$. 10) $\bar{K}_{c_2}$ is weakly decreasing in $|R|$.*

Proof. From Section A.1.1 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 (6)

$$\Psi(x_i^*; R, A_i) = \int_{R-\Delta_i}^{R} (R - \tau)\, d\tau + \int_{R}^{R+\Delta_i} (\tau - R)\, d\tau = \Delta_i^2$$

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

$$\Psi(x_i^*; R, A_i) = \int_{-1}^{R} (R - \tau)\, d\tau + \int_{R}^{R+\Delta_i} (\tau - R)\, d\tau = \frac{1}{2}(1 + R)^2 + \frac{1}{2}\Delta_i^2.$$

Hence we get

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

noting that $\Psi\left(x_i^*;R,A_i\right)$ might equal 1 only in the middle range (unless $R=0$), and in particular when $1=\frac{1}{2}\left(1+R\right)^2+\frac{1}{2}\Delta_i^2\Leftrightarrow 1-R^2-2R=\Delta_i^2$ we get by (7) that

$$A_{i+1}\equiv f\left(R,A_i\right)=\begin{cases}1-\Delta_i^2 & \text{when } 0\leq\Delta_i\leq 1+R\\ 1-\left(\frac{1}{2}\left(1+R\right)^2+\frac{1}{2}\Delta_i^2\right) & \text{when } 1+R<\Delta_i<\sqrt{1-R^2-2R}\\ 0 & \text{when } \sqrt{1-R^2-2R}\leq\Delta_i\end{cases}. \tag{19}$$

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

$$\Delta_i=\left(\bar{K}A_i\right), \tag{20}$$

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 4 shows and will now be proved, the two regions in which $A_{i+1}\neq 0$ are concave. Using (19) and (20) we get

$$\frac{df}{dA_i}=\left\{\begin{array}{ll}-\frac{2}{\alpha-\beta}\Delta_i^2A_i^{-1} & \text{when } \Delta_i\leq 1+R\\ -\frac{1}{\alpha-\beta}\Delta_i^2A_i^{-1} & \text{when } 1+R<\Delta_i<\sqrt{1-R^2-2R}\end{array}\right\}>0$$

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

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

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

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

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

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 (20) that Δ_i is strictly positive, hence, by (19), $A_{i+1} < 1$, 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

$$\bar{K} = \frac{(1+R)^{\alpha-\beta}}{1-(1+R)^2} \tag{21}$$

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

$$1-(1+R)^2 = (1+R)^{\alpha-\beta}/\bar{K} \Rightarrow \{\text{using (19) and (20)}\} \Rightarrow A_{i+1}|_{\Delta_i=1+R} = A_i|_{\Delta_i=1+R},$$

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 (21)) by $\bar{K}_{c_2}$. Thus follows part (9), and part (10) follows from the fact that f increases in $\bar{K}$ and $|R|$ (by parts (5) and (6)). ■

Proof of Proposition 3 Part (1) follows from Lemma 3 parts (9) and (10). Part (2): From definition 1 it follows that any convergence to a lower steady state constitutes a revolution because the approval falls over several periods. Hence 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 3), and in particular, if the shock is such that $\bar{K}$ goes below $\bar{K}_{c_2}$, then A_i converges to zero and the regime completely falls (part (9) of Lemma 3). 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 3), 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 ($x(t) = t$).

A.5 Proof of Proposition 4

We first outline some helpful analytical results and then prove the actual proposition. When $\alpha > 1$, $\beta \geq 1$, every type $t > R$ has a unique inner solution $x^*(t) \in]R, t[$ and every type $t < R$ has a unique inner solution $x^*(t) \in]t, R[$, with this solution being determined by equation (15) (see Section A.1.4). 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 $\sigma \equiv |x^*(t) - R|$ and $\tau \equiv |t - R|$ yields

$$\begin{aligned} K_i \beta \sigma^{\beta-1} &= \alpha (\tau - \sigma)^{\alpha-1} \\ &\Leftrightarrow \tau = \sigma + \left(\frac{K_i \beta}{\alpha}\right)^{\frac{1}{\alpha-1}} \sigma^{\frac{\beta-1}{\alpha-1}}. \end{aligned} \tag{22}$$

We turn now to calculating $\Psi(x_i^*; R, A_i)$. To do that, we first remind that $\Psi(x_i^*; R, A_i)$ is the sum of deviations from R (i.e. the sum of $\sigma(t)$ over all 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 (22). 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).[34] Finally, using the symmetry of $\sigma(\tau)$ around 0 we can

[34]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 $\sigma(1 + R)$.

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

$$\Psi(x_i^*;R,A_i)=\int_0^{1+R}\sigma(\tau)\,d\tau+\int_0^{1-R}\sigma(\tau)\,d\tau$$
$$=\int_0^{\check{\sigma}\equiv\sigma(1+R)}[(1+R)-\tau(\sigma)]\,d\sigma+\int_0^{\hat{\sigma}\equiv\sigma(1-R)}[(1-R)-\tau(\sigma)]\,d\sigma$$
$$=\int_0^{\check{\sigma}\equiv\sigma(1+R)}\left[(1+R)-\sigma-\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}\right]d\sigma+\int_0^{\hat{\sigma}\equiv\sigma(1-R)}\left[(1-R)-\sigma-\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}\right]d\sigma$$

$$=(1+R)\,\check{\sigma}-\frac{\check{\sigma}^2}{2}+(1-R)\,\hat{\sigma}-\frac{\hat{\sigma}^2}{2}-\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}. \tag{23}$$

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

Lemma 4. *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 $\bar{K}_{c_3}$ such that a stable steady state with a regime and $A_{ss}>0$ exists iff $\bar{K}>\bar{K}_{c_3}$. 8)$\bar{K}_{c_3}$ 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 (7) $A=\max\{0,1-\Psi(x_i^*;R,A_i)\}$ and note that

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

Using $\check{\sigma}$ and $\hat{\sigma}$ in the FOC in (11) we get

$$\alpha\left(1+R-\check{\sigma}\right)^{\alpha-1}=\bar{K}A_i\beta\check{\sigma}^{\beta-1} \tag{24}$$

$$\alpha\left(1-R-\hat{\sigma}\right)^{\alpha-1}=\bar{K}A_i\beta\hat{\sigma}^{\beta-1}. \tag{25}$$

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 (7) 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 $\Psi\left(x_i^*;R,A_i\right)=\frac{(1-R)^2+(1+R)^2}{2}\geq 1$, with equality only for $R=0$. From (6) and (7) 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:

$$\begin{aligned}\frac{d^2\Psi\left(\sigma_i;R,A_i\right)}{dA_i^2}= & -\frac{1}{\alpha-1}\left(\frac{1}{\alpha-1}-1\right)A_i^{\frac{1}{\alpha-1}-2}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}\\ & -\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-1}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\left(\check{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\check{\sigma}}{dA_i}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\hat{\sigma}}{dA_i}\right).\end{aligned} \tag{26}$$

Note that $\frac{d\check{\sigma}}{dA_i}$ and $\frac{d\hat{\sigma}}{dA_i}$ are both negative.[35] 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. We now investigate the case $1<\alpha<2$. Revisiting equation (22) we can write

$$\begin{aligned}H=\sigma+\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}-\tau=0\Rightarrow\frac{d\sigma}{dA_i}=-\frac{\frac{dH}{dA_i}}{\frac{dH}{d\sigma}}=-\frac{\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-1}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}}{1+\frac{\beta-1}{\alpha-1}A_i^{\frac{1}{\alpha-1}}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}-1}}\\ =\{\text{using (22)}\}=-\frac{\frac{1}{\alpha-1}A_i^{-1}\left(\tau-\sigma\right)}{1+\frac{\beta-1}{\alpha-1}\left(\tau-\sigma\right)\sigma^{-1}}.\end{aligned} \tag{27}$$

Rewriting (26)

$$\frac{d^2\Psi\left(\sigma_i;R,A_i\right)}{dA_i^2}=-\frac{1}{\alpha-1}A_i^{\frac{1}{\alpha-1}-2}\left(\frac{\bar{K}\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\times$$

[35] 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.

$$\left[\left(\frac{1}{\alpha-1}-1\right)\frac{\check{\sigma}^{\frac{\beta-1}{\alpha-1}+1}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}+1}}{\frac{\beta-1}{\alpha-1}+1}+A_i\left(\check{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\check{\sigma}}{dA_i}+\hat{\sigma}^{\frac{\beta-1}{\alpha-1}}\frac{d\hat{\sigma}}{dA_i}\right)\right]$$

Using the FOC $\left(\frac{K_i\beta}{\alpha}\right)^{\frac{1}{\alpha-1}}\sigma^{\frac{\beta-1}{\alpha-1}}=\tau-\sigma$ and (27) we get

$$\frac{d^2\Psi\left(\sigma_i;R,A_i\right)}{dA_i^2}=\begin{array}{l}-\frac{1}{\alpha-1}A_i^{-2}\left[\sigma\left(\tau-\sigma\right)\left(\frac{1}{\alpha-1}-1\right)\frac{1}{\frac{\beta-1}{\alpha-1}+1}-\left(\tau-\sigma\right)\frac{\frac{1}{\alpha-1}(\tau-\sigma)}{1+\frac{\beta-1}{\alpha-1}\frac{\tau-\sigma}{\sigma}}\right]\Big|_{\tau=1+R}\\ -\frac{1}{\alpha-1}A_i^{-2}\left[\sigma\left(\tau-\sigma\right)\left(\frac{1}{\alpha-1}-1\right)\frac{1}{\frac{\beta-1}{\alpha-1}+1}-\left(\tau-\sigma\right)\frac{\frac{1}{\alpha-1}(\tau-\sigma)}{1+\frac{\beta-1}{\alpha-1}\frac{\tau-\sigma}{\sigma}}\right]\Big|_{\tau=1-R}\end{array}. \tag{28}$$

Note that

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

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 (28) 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 $A_{i+1}=\max\left\{0,1-\Psi_i\left(\sigma_i;R,A_i\right)\right\}$, we get that $A_{i+1}\left(A_i\right)$ is first convex then concave, or convex throughout, or concave throughout, which proves part (3). Differentiating equation (23) w.r.t. R and then using (24) and (25) yields

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

(by the monotonicity of $\sigma\left(\tau\right)$), implying that A_{i+1} decreases in $|R|$, which proves part (4). Next, differentiating equation (23) by $\bar{K}$ and then using (24) and (25) yields

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

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

$A_{i+1} = f(1)$ never reaches 1. This further implies, together with the fact that the phase diagram $A_{i+1} = f(A_i)$ – see Figure 6 – 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 (22) is smaller than 1/2. The strict monotonicity of $\sigma(\tau)$ implies then that the total sum of deviations from the regime ($\Psi(x_i^*; R, A_i)$) 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.[36] Furthermore, this happens for finite $\bar{K}$. Together with this result, part (5) implies that $f(A_i)$ 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 $\bar{K}_{c_3}$. Thus follows part (7), and part (8) follows from the fact that $f(A_i)$ 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).[37] ■

Proof of Proposition 4 Part (1) follows from Lemma 4 parts (7) and (8). Part (2): From definition 1 it follows that any convergence to a lower steady state constitutes a revolution because the approval falls over several periods. Hence 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 $|A_{ss} - A_{uss}|$ (when A_{uss} exists), would result in a revolution.

[36] If $R = 0$ and $A_{i+1} = f(1/2) > 1/2$, 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.

[37] 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 $A_{i+1} = f(A_i)$ results as well in a decrease of A_{ss}.

A negative shock to the force of the regime reduces A_{i+1} (part (5) of Lemma 4), and in particular, if the shock is such that $\bar{K}$ goes below $\bar{K}_{c_3}$, A converges to zero over time and the regime completely falls (part (7) of Lemma 4). 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 4), and in particular a revolution would start if the approval decreases sufficiently to eliminate the per-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 will be two-sided. Part (4): Follows from part (3) of Proposition 1.

A.6 Proofs of optimal β

We start by proving a few helpful results.

Lemma 5. *1) If $K \geq 1$ then, for any $\alpha > 0$, $\beta^* < 1$ and $x(t) = 0 \, \forall t$.*

Proof. $\alpha \leq 1$: Suppose $\beta \leq 1$. From Section A.1.1 we know that all individuals have a corner solution, namely either $x = 0$ or $x = t$. The loss when choosing $x = 0$ is t, while the loss when choosing $x = t$ is Kt^β. When $K > 1$, we get that $Kt^\beta > t^\beta \geq t$ for any $t \leq 1$ and $\beta \leq 1$, implying that $x(1/2) = x(1) = 0$, i.e. the dissent is minimal.

$\alpha > 1$: Let $\beta \to 0$. Then for any $x > 0$, $L(x;t) = K + D(x;t)$ which has a min point for $x(t) = t$ where $L(x;t) = K$. For $x = 0$ $L(x,t) = t^\alpha < K \; \forall t \leq 1$. Hence $x^* = 0 \; \forall t$.

Thus, for any α, the regime cannot do better by choosing $\beta > 1$, and in fact would do strictly worse by doing so, because $\beta > 1$ implies that $S'(0) = 0$, hence no individual would choose $x = 0$. ■

Lemma 6. *If $\hat{\beta}$ minimizes the dissent of t, and the chosen stance is such that $x^*(t) \notin \{0, t\}$, then $1 + \hat{\beta} ln(x^*(t)) = 0$.*

Proof. $x^*(t) \notin \{0, t\}$ implies that we can apply the implicit function theorem to equation (10) using the fact that $L'(x^*(t);t) = 0$, which yields (for $R = 0$)

$$\frac{dx^*}{d\beta} = -\frac{Kx^{*\beta-1}(1 + \beta ln(x^*))}{(\alpha - 1)\,\alpha\,(t - x^*)^{\alpha-2} + (\beta - 1)\,\beta K x^{*\beta-2}}. \tag{29}$$

Since $\hat{\beta}$ minimizes the dissent of t, it follows that $\frac{dx^*}{d\beta}|_{\hat{\beta}} = 0$, implying that $1 + \hat{\beta} ln(x^*(t)) = 0$. ■

Lemma 7. *The function $L(x;1)$ has an inner local min point at x_0 if and only if there exists $x_0 \in]0,1[$ such that*

$$K = K^* \equiv \frac{\alpha(1-x_0)^{\alpha-1}}{\beta x_0^{\beta-1}}. \tag{30}$$

Proof. In an inner local min point x_0, $L'(x_0) = D' + KP' = 0$. Isolating K and substituting $t = 1$, we get that an inner local min point exists if and only if there exists $x_0 \in]0,1[$ such that

$$K = K^* \equiv \frac{\alpha(1-x_0)^{\alpha-1}}{\beta x_0^{\beta-1}}.$$

■

A.6.1 Proof of Lemma 1

If $\bar{K} > 1$ and $A_i = 1$ then $K_{i+1} > 1$ so that by Lemma 5, $\beta^*_{i+1} < 1$ and $x_{i+1}(t) = 0\ \forall t$. Then $A_{i+1} = 1 = A_i$, hence this is a steady state. It is stable, since there exists a neighborhood of $A_i = 1$, where $A_{i+1} = 1$. This proves Lemma 1.

A.6.2 Proof of part 1 of Proposition 5

Lemmas 5 and 1 show that a steady state exists when $K_i > 1$. A negative shock to K (or to A that consequently affects K) of a particular size will lead to $K_i < 1$. Given our definition of revolution as a sequence of periods where aggregate dissent increases (Definition 3), to prove statement 1 of the proposition it is sufficient to show that there exists (a range of) $K_i < 1$ such that $\beta^*_i < \alpha$ and $x(1) > 0$ and $x(1/2) = 0$.[38] Lemma 10 below proves that such K_i exist for $\alpha \in]\underline{\alpha}, 1[$ (where $\underline{\alpha} \approx 0.53$).

Lemma 8. *If $K < K^{**} \equiv \alpha ln(2) 2^{1/ln(2)-\alpha}$ and $\alpha < 1 < \beta$, then $x^*(1) > 1/2$.*

Proof. When $\beta > 1$, $x^*(t) \neq 0\ \forall t$. Then either $x^*(1) = 1$ (a corner solution), implying immediately that $x^*(1) > 1/2$; or $x^*(1) \in]0,1[$. In this case, larger K implies smaller $x^*(1)$, and we look for the value of K below which $x^*(1) > 1/2$. Substituting $x^*(1) = 1/2$ in equation (30) yields $K^*(x_0 = \frac{1}{2}) = \frac{\alpha}{\beta} 2^{\beta-\alpha}$. To make sure that $x^*(1) > 1/2$ for any $\alpha < 1 < \beta$, K has to be smaller than $K^*(x_0 = \frac{1}{2})$ for any $\beta > 1$. From Lemma 6 we get that the value of $\beta > 1$ such that $\frac{dx^*}{d\beta} = 0$ and $x^*(1) = 1/2$ is $\beta = 1/ln(2)$, implying that K has to be smaller than $K^*(x_0 = \frac{1}{2}, \beta = 1/ln(2)) = K^{**} = \alpha ln(2) 2^{1/ln(2)-\alpha}$. ■

[38] Note that it is always possible to find a $\bar{K}$ such that this $K_i < 1$ leads to increased dissent in the next period.

Lemma 9. $(1/2)^\alpha < K^{**}$ *if and only if* $\alpha > \underline{\alpha}$, *where* $\underline{\alpha} \equiv \frac{1}{ln(2)2^{1/ln(2)}} \approx 0.53$.

Proof. $(1/2)^\alpha < K^{**} = \alpha ln(2) 2^{1/ln(2)-\alpha} \Longleftrightarrow 1 < \alpha ln(2) 2^{1/ln(2)} \Longleftrightarrow \alpha > \underline{\alpha} = \frac{1}{ln(2)2^{1/ln(2)}}$. ■

Lemma 10. *For any* $\alpha \in]\underline{\alpha}, 1[$ *there exists a range of* K *such that* $\beta^*(K) < \alpha$ *and* $x^*(1/2) = 0$ *while* $x^*(1) = 1$.

Proof. Here we use the notation that μ is the share of moderates in society. The range of K to which the lemma applies is $K \in](1/2)^\alpha, K^{**}[$. By Lemma 9, this range is non empty if and only if $\alpha > \underline{\alpha}$. Since $\alpha \in]\underline{\alpha}, 1[$, we know from Lemma 8 that for any α in this range and any $\beta > 1$, we have $x^*(1) > 1/2$. This further implies that $t = 1/2$ does not have an inner solution when $\beta > 1$, because we know (from Appendix A.1.3) that, whenever $\alpha < 1 < \beta$, if $t<t'$ both have inner solutions then $x(t) > x(t')$, implying that for $x^*(1/2)$ to be an inner solution it must be larger than $x^*(1) > 1/2$. Thus, if the regime chooses $\beta > 1$, $x^*(1/2) = 1/2$ (this is the only possible corner solution when $\alpha < 1 < \beta$), and we get that the total dissent is strictly larger than $\frac{1}{2}[\mu + (1-\mu)] = \frac{1}{2}$. In comparison, by choosing $\beta \leq 1$, the regime would force everyone to choose a corner solution (see Section A.1.1). Since $K < K^{**}$ and $K^{**} < 1$ (for any $\alpha < 1$), the corner solution for $t = 1$ will always be $x^*(1) = 1$. As for the stance chosen by $t = 1/2$, the regime aims it to be $x^*(1/2) = 0$. A necessary condition for this to be the case is that $\beta^*(K) < \alpha$ (as otherwise $L(1/2; 1/2) = K(1/2)^\beta < (1/2)^\beta < (1/2)^\alpha = L(0, 1/2)$). Depending on the exact value of $\beta < 1$ that is chosen, this may or may not suffice to get $t = 1/2$ to choose $x^*(1/2) = 0$. However, by taking $\beta \to 0$, the regime can assure that this is the case for any $K > (1/2)^\alpha$. Overall, this yields $x^*(1) = 1$ and $x^*(1/2) = 0$, resulting in a total dissent of $1 - \mu$. For $\mu = 1/2$ this implies less dissent than under any $\beta > 1$, hence this is the optimal choice of the regime. ■

A.6.3 Proof to part 2 of Proposition 5

Lemma 5 and 1 show that a steady state exists when $K_i > 1$. A negative shock to K (or to A that then affects K) of a particular size will lead to $K_i < 1$. Given our definition of revolution as a sequence of periods where aggregate dissent increases (Definition 3), to prove statement 2 of the proposition, it is sufficient to show that there exists (a range of) $K_i < 1$ such that $\beta^*(K_i) > \alpha$ (Lemma12) and that if $\alpha > \tilde{\alpha}$ then there exists (a range of) $K_i < 1$ such that $\beta^*(K_i) > \alpha$ and moderates dissent strictly more than extremists

(Lemma 13).[39]

Lemma 11. *For any $\alpha < 1 < \beta$, there exists a K_{min} such that $L(x;1)$ has an inner local min point x_0 for all $K > K_{min}$. $lim_{\beta\to\infty} K_{min} = 0 < (1/2)^{\alpha}$.*

Proof. In an inner local min point x_0, $K = K^*$ as given in equation (30). Note that K^* is continuous, positive and goes to infinity at $x_0 \to 0$ and 1. Thus, K as a function of x_0 must have a minimum. This proves the first part of the lemma.

For any $x_0 \in]0,1[$ there is a K that satisfies (30). The question is whether this value of K satisfies $K \in]0,(1/2)^{\alpha}[$. Let us find this minimum $K_{min} \equiv min_{x_0\in]0,1[} K^*$. $\frac{dK^*}{dx_0} = 0 \Rightarrow \left[-\frac{\alpha-1}{1-x_0} + \frac{1-\beta}{x_0}\right] K^* = 0 \Rightarrow ... \Rightarrow (\beta-\alpha)x_0 = (\beta-1)$. Hence, x_0 corresponding to K_{min} equals $\frac{\beta-1}{\beta-\alpha} < 1$. Plugging this x_0 into (30) gives

$$K_{min} = \frac{\alpha}{\beta}(1-\alpha)^{\alpha-1}(\beta-1)^{1-\beta}(\beta-\alpha)^{\beta-\alpha}$$

where $lim_{\beta\to\infty} K_{min} = \alpha(1-\alpha)^{\alpha-1} lim_{\beta\to\infty}\beta^{-\alpha} = 0 < (1/2)^{\alpha}$ which proves the second sentence. ■

Lemma 12. *For any $\alpha < 1$ there exists a range of K such that $\beta^*(K) > 1$.*

Proof. Let $K < (1/2)^{\alpha}$. Suppose $\beta \leq 1$. Then (see Appendix A.1.1) $x(t) \in \{0,t\}$. Comparing these corner solutions $x^*(t) = t$ if and only if $K < (t)^{\alpha-\beta}$ which is true since $K < (1/2)^{\alpha} \leq (t)^{\alpha-\beta}$ when $\beta > 0$ and $t \in \{1/2, 1\}$. Hence, $\beta < 1$ yields full dissent and any $\beta > 1$ that yields $x(t) < t$ for some t is preferred by the regime. We now show that there exist $\beta > 1$ such that $x(1) < 1$.

Consider $t = 1$. When $\alpha < 1 < \beta$ we know (from Appendix A.1.3) that there may exist two local min points for $x \in [0,1]$: a corner min $x(t) = t$ and an interior min where $x(t) \in]0,t[$ in which the FOC holds. From Lemma 11 we know that for sufficiently large β the inner local min point exists for any $K > 0$. We now show that the inner local min point is the global min point for sufficiently large β. $\lim_{\beta\to\infty} P' = \lim_{\beta\to\infty} \frac{x^{\beta-1}}{1/\beta} =$ {using L'Hopital's rule}$= \frac{x^{-1}x^{\beta}lnx}{-1/\beta^2} = -\lim_{\beta\to\infty} P'\beta lnx$. This can only hold if $P' = 0$ or if $\lim_{\beta\to\infty} \beta lnx = -1$. In the inner solution, which we have shown exists, $P' = -D' = \alpha(1-x)^{\alpha-1} > 0 \forall x \in [0,1]$. Hence, in the inner min point $G \equiv \lim_{\beta\to\infty} lnx^{\beta} = -1$ has to hold. Furthermore, since $\lim_{\beta\to\infty} P = 0$ $\forall x \in [0,1[$ it has to be that the inner min point is at $x_0 \to 1$. Now note that $G = \lim_{\beta\to\infty} lnP(x_0) = -1 \to \lim_{\beta\to\infty} P(x_0) =$

[39]Note that it is always possible to find a $\bar{K}$ such that this $K_i < 1$ leads to increased dissent in the next period.

$e^G = e^{-1}$. The inner min point is preferred over the corner at $x = 1$ if and only if $\triangle L \equiv L(x_0, 1) - L(1, 1) = KP(x_0) + D(t - x_0) - K$. At $x_0 \to 1$ $D = 0$ and $P = 1/e < 1$ hence $\triangle L < 0$. Hence, there exists a sufficiently large β such that $x(1) < 1$ implying that $\beta^* > 1$. ■

Lemma 13. *If $1 > \alpha > \tilde{\alpha} \approx 0.215$, then there exists a range of K such that, under the optimal sanctioning $\beta^*(K) > 1$ and moderates dissent strictly more than extremists.*

Proof. **We first prove that when $K \nearrow 1$ and $1 > \alpha > \tilde{\alpha} \approx 0.215$, we have $\beta^*(K) > 1$:**

First, it can easily be shown that when $K \in](1/2)^\alpha, 1[$ and $\beta \leq 1$ then $x^*(1) = 1$ and $x^*(1/2) = 0$ (see also the proof of Lemma 10). Hence, a sufficient condition for $\beta^*(K) > 1$ is that $x^*(1) < 1/2$ for some combination of $\beta > 1$ and $K \in](1/2)^\alpha, 1[$. When $\alpha \in]0, 1[$ and $\beta > 1$, we know (from Appendix A.1.3) that there are two candidate min points: a corner solution $x(t) = t$ and an interior solution where $x_0(t) \in]0, t[$ in which the FOC holds. Hence, a sufficient condition for $\beta^* > 1$ is that, for $t = 1$, the interior min point is the global min and that, in this interior min point, $x_0(1) < 1/2$. This is true if A) $L(1/2; 1, K) \leq L(1; 1, K)$ and B) $L'(x, 1, K)_{x=1/2} > 0$.[40] We will now show that there exists $\beta > 1$ such that (A) and (B) hold for $K \nearrow 1$. The candidate β we show this for is $\beta = 1/\alpha > 1$. Note that $\beta = 1/\alpha$ may not be optimal, but since we show it does better than any $\beta \leq 1$, β^* necessarily is greater than 1.

A) When $K \nearrow 1$ and $\beta = 1/\alpha$, then $L(1/2, 1, K) \leq L(1, 1, K) \leftrightarrow$

$$(1/2)^{1/\alpha} + (1/2)^\alpha \leq 1. \tag{31}$$

Note that the inequality holds only weakly in the limits, i.e., when $\alpha \to 0$ and $\alpha \to 1$. For $\alpha = 1/2$ the inequality holds strictly, implying that the LHS has at least one local min point. Differentiating LHS and writing down the FOC we get

$$\frac{dLHS}{d\alpha} = -\frac{1}{\alpha^2}(1/2)^{1/\alpha} ln(1/2) + (1/2)^\alpha ln(1/2) = -ln(2)[-\frac{1}{\alpha^2}(1/2)^{1/\alpha} + (1/2)^\alpha] = 0. \tag{32}$$

The FOC in (32) holds when

$$-\frac{1}{\alpha^2}(1/2)^{1/\alpha} + (1/2)^\alpha = 0 \leftrightarrow (1/2)^{1/\alpha} = \alpha^2 (1/2)^\alpha.$$

[40] Note that (B) implies that the inner min point $x_0 < 1/2$ since L, when $\beta > 1$, is a first decreasing, then increasing and then decreasing function of $x \in [0, 1]$. The inner min point is located at the first place where $L' = 0$, i.e. before the increasing part of L.

Plugging it in the LHS of the inequality condition (31) we get:

$$(1/2)^{\alpha}(\alpha^2+1) \leq 1.$$

This inequality, where we have constrained α to obey the FOC, is weakly true at the corners ($\alpha = 0$ and $= 1$). Differentiating the LHS we get

$$\frac{d}{d\alpha}(1/2)^{\alpha}(\alpha^2+1) = (1/2)^{\alpha} ln(1/2)(\alpha^2+1) + (1/2)^{\alpha} 2\alpha$$

$$= (1/2)^{\alpha}(ln(1/2)(\alpha^2+1) + 2\alpha).$$

This expression equals 0 when

$$ln(2) = \frac{2\alpha}{(\alpha^2+1)},$$

where LHS>RHS when $\alpha \to 0$ and LHS<RHS when $\alpha \to 1$ and the RHS is increasing for any $\alpha < 1$. Thus, $(1/2)^{\alpha}(\alpha^2+1)$ has one extremum in $\alpha \in]0,1[$. If this extremum would be a max point, implying that $(1/2)^{\alpha}(\alpha^2+1) \geq 1$ for any $\alpha \in]0,1[$, it would have also implied that (31) does not hold at any extremum point of the LHS of (31), contradicting the fact that this LHS has at least one local min point. Thus, we conclude that the extremum of $(1/2)^{\alpha}(\alpha^2+1)$ is a min point, implying further that $(1/2)^{\alpha}(\alpha^2+1) \leq 1$ for any $\alpha \in [0,1]$, so that any extremum point of the LHS of (31) has to be a min point, implying further that the extremum is unique. Thus, $(1/2)^{1/\alpha} + (1/2)^{\alpha} \leq 1$ holds for all $\alpha < 1$.

B) When $K \nearrow 1$ and $\beta = 1/\alpha$, condition (B) becomes $L'(x,1,K)_{x=1/2} > 0 \leftrightarrow P'(x,1,K)_{x=1/2} > D'(x,1,K)_{x=1/2} \leftrightarrow \frac{1}{\alpha}(1/2)^{1/\alpha-1} > \alpha(1/2)^{\alpha-1}$ which (taking logs and rewriting) is

$$ln(1/2)/2 > \alpha ln(\alpha)/(1-\alpha^2). \tag{33}$$

The LHS of this inequality is roughly equal to -0.34. As for the RHS, using L'Hopital's rule we get

$$lim_{\alpha \searrow 0} ln(\alpha)/((1-\alpha^2)/\alpha) = lim_{\alpha \searrow 0} \frac{1/\alpha}{-1-1/\alpha^2} = lim_{\alpha \searrow 0} -\alpha = 0 > -0.34$$

and

$$lim_{\alpha \nearrow 1} \alpha ln(\alpha)/(1-\alpha^2) = lim_{\alpha \nearrow 1} \frac{ln(\alpha)+1}{-2\alpha} = -1/2 < -0.34.$$

Hence (33) does not hold at $\alpha \searrow 0$ but holds at $\alpha \nearrow 1$. We now verify that there is

only one cutoff value $\tilde{\alpha}$ where the RHS intersects the $LHS = -0.34$. We do this by showing that the RHS has no inner extrema, which implies – given that it is continuously differentiable in the range $\alpha \in]0,1[$ – that it must be strictly monotonous in this range. Differentiating the RHS of (33) to find an inner solution we get

$$\frac{dRHS}{d\alpha} = 1/(1-\alpha^2) + ln(\alpha)\frac{(1-\alpha^2)+2\alpha^2}{(1-\alpha^2)^2} = 0 \leftrightarrow 1/(1+\alpha^2) = -ln(\alpha)/(1-\alpha^2) \leftrightarrow$$

$$(1-\alpha^2)/(1+\alpha^2) = -ln(\alpha). \tag{34}$$

The LHS of (34) monotonically decreases from 1 to 0 as α goes from 0 to 1, while the RHS is a convex decreasing function, going from ∞ to 0 as α goes from 0 to 1. The equation thus holds at $\alpha = 1$. For (34) to have another solution (i.e., an inner solution), these two functions need to intersect at some $\alpha \in]0,1[$. Since both functions are continuously differentiable in the range $\alpha \in]0,1[$, a necessary condition for intersection is that they would have the same slope for at least one $\alpha \in]0,1[$ (a value in-between the two intersection points). Differentiating both sides of (34) and equating, we get

$$-2\alpha\frac{2}{(1+\alpha^2)^2} = -1/\alpha \leftrightarrow 4\alpha^2 = (1+\alpha^2)^2 \leftrightarrow$$

$$0 = 1 - 2\alpha^2 + \alpha^4 = (1-\alpha^2)^2$$

which holds only when $\alpha = 1$. Hence (34) has only one solution (at $\alpha = 1$) implying the RHS of (33) has no inner extrema, which implies that it must be monotonous in the range $\alpha \in [0,1]$. This fact, along with (33) not holding for $\alpha \to 0$ and holding for $\alpha \to 1$, implies that condition (B) holds for any α larger than some cutoff $\tilde{\alpha}$ defined by $ln(1/2)/2 = \alpha ln(\alpha)/(1-\alpha^2)$. Solving this equation numerically yields $\tilde{\alpha} \approx 0.215$.

We have thus verified that, for any $\alpha \in]\tilde{\alpha}, 1[$, there exists a value of $\beta > 1$ s.t. $x(1) < 1/2$ for $K \nearrow 1$. Thus, by continuity, this holds also for values of K sufficiently close to 1.

We next prove that in that same range of $K \nearrow 1$ and for $1 > \alpha > \tilde{\alpha} \approx 0.215$ and $\beta^*(K) > 1$, moderates dissent strictly more than extremists: This is true if $x^*(1) < x^*(1/2)$. Suppose by negation that $x^*(1) \geq x^*(1/2)$. Having shown already that $x^*(1) < 1/2$, this implies that $x^*(1/2)$ must be smaller than 1/2, i.e. $t = 1/2$ must have an inner solution to the optimization problem, just like $t = 1$ does. However, we know (from Appendix A.1.3) that, whenever $\alpha < 1 < \beta$, if $t<t'$ both have inner solutions then

$x(t) > x(t')$, in contradiction to the assumption that $x^*(1) \geq x^*(1/2)$. ■

A.6.4 Proof to part 3 of Proposition 5

Lemma 5 and 1 show that a steady state exists when $K_i > 1$. A negative shock to K (or to A that then affects K) of a particular size will lead to $K_i < 1$. Given our definition of revolution as a sequence of periods where aggregate dissent increases (Definition 3), to prove statement 3 of the proposition, it is thus sufficient to show that there exists (a range of) $K_i < 1$ such that $\beta^*(K_i) > 1\}$ and moderates dissent strictly less than extremists.[41] Lemma 14 below proves that $\beta^*(K) > 1$, and the fact that, in this case, extremists dissent strictly more than moderates follows from Section A.1.4.

Lemma 14. *For any $\alpha > 1$, there exists a range of K such that $\beta^*(K) > 1$.*

Proof. Let K' be defined implicitly by $K_\alpha = \left(\frac{K}{\alpha}\right)^{\frac{1}{\alpha-1}} = 1/2 - 1/e$ and let $K^{***} \equiv min\{K', (1/2)^\alpha\}$. The range of K for which this lemma holds is $K < K^{***}$, where we note that $K^{***} > 0$ for any $\alpha > 1$. From Section A.1.4 we know that, when $\alpha > 1$, if also $\beta > 1$ then individuals have an inner solution to the optimization problem. If $\beta \searrow 1$, we get from (15) that $x^*(t)$ in the inner solution is such that $\alpha(t-x^*)^{\alpha-1} = K \Rightarrow t - x^* = K_\alpha < 1/2 - 1/e$ (and therefore $K < K'$). This implies that both $x^*(1/2)$ and $x^*(1)$ are larger than $1/e$, which further implies (see (29)) that $1 + \beta ln(x^*)$ is positive hence $\frac{dx^*}{d\beta}|_{\beta=1} < 0$. This means that by increasing β (locally) beyond 1, the regime decreases dissent hence improves its approval. As β increases further (and x^* decreases further), the expression $1 + \beta ln(x^*)$ decreases monotonically towards 0, at which point $\frac{dx^*}{d\beta}$ switches signs from negative to positive. For a large enough β, $\frac{dx^*}{d\beta} > 0$ for both $x^*(1/2)$ and $x^*(1)$, implying that $\beta^*(K)$ is finite. Turning our attention now to $\beta < 1$, we first remind the reader that when $\beta < 1 < \alpha$, the only potential corner solution is $x^*(t) = 0$ (see Section A.1.2). This corner solution will be chosen by neither type here, since it is dominated even by the other (not-chosen) corner solution of $x^*(t) = t$ for both $t = 1/2$ (since $L(1/2; 1/2) = K(1/2)^\beta < K < (1/2)^\alpha = L(0; 1/2)$) and $t = 1$ (since $L(1; 1) = K < (1/2)^\alpha < 1 = L(0; 1)$). Thus, we have to consider only the inner solutions for the case of $\beta < 1$. Since $\frac{dx^*}{d\beta}|_{\beta=1} < 0$, decreasing β slightly below 1 decreases the approval of the regime. As β decreases further (and x^* increases further), the expression $1 + \beta ln(x^*)$ increases monotonically hence stays positive, implying that

[41] Note that it is always possible to find a $\bar{K}$ such that this $K_i < 1$ leads to increased dissent in the next period.

$\frac{dx^*}{d\beta}$ stays negative. Thus, any $\beta < 1$ is dominated by $\beta = 1$, which is itself dominated by the optimal (and finite) $\beta^*(K) > 1$. ■

A.6.5 Proof of Proposition 6

Lemma 15. *Let $R = 0$, $\alpha = 1$, and suppose that $K \leq 1/2$ and all individuals have either $t = 1/2$ or $t = 1$. Then, for any $\mu \in [0, 1]$, the optimal β of the regime is such that $\beta > 1$.*

Proof. First suppose that the regime chooses $\beta \leq 1$. From Section A.1.1 we know that all individuals have a corner solution, namely either $x = 0$ or $x = t$. The loss when choosing $x = 0$ is t, while the loss when choosing $x = t$ is Kt^β. When $K \leq 1/2$, we get that $Kt^\beta \leq K \leq t$ for any $t \geq 1/2$ (with at least one of the inequalities being strict), implying that $x(1/2) = 1/2$ and $x(1) = 1$, i.e. the dissent is maximal. If instead the regime chooses $\beta > 1/K > 1$,then $x(1) < 1$ because $L'(1) = -1 + K\beta > 0$, implying that an individual with $t = 1$ has a profitable deviation from choosing $x = t$. Since the regime seeks to minimize dissent, any $\beta > 1/K$ dominates any $\beta \leq 1$, implying further that the optimal β of the regime is such that $\beta > 1$. ■

Lemma 16. *Let $R = 0$, $\alpha = 1$, and suppose all individuals have either $t = 1/2$ or $t = 1$. If, for any $\mu \in [0, 1]$, the optimal β of the regime is such that $\beta > 1$, then it is independent of μ.*

Proof. The case of $1 = \alpha < \beta$ is analyzed at the end of Section A.1.4. We show there that types sufficiently close to the regime choose $x(t) = t$, while types sufficiently far from the regime choose the same inner solution x s.t. $S'(|x - R|) = 1$ $(= D')$. Denote this inner solution by $\tilde{x}$. Then it immediately follows that, regardless of the vale of μ, the regime will choose the value of β that minimizes $\tilde{x}$, and this value is itself independent of μ because it is determined by the equation $S'(|x - R|) = 1$, which does not contain μ. ■

Lemma 17. *Let $R = 0$, $\alpha = 1$, and suppose that $K > 1$ and all individuals have either $t = 1/2$ or $t = 1$. Then, for any $\mu \in [0, 1]$, the regime minimizes dissent by choosing any $\beta \leq 1$.*

Proof. Suppose that the regime chooses $\beta \leq 1$. From Section A.1.1 we know that all individuals have a corner solution, namely either $x = 0$ or $x = t$. The loss when choosing $x = 0$ is t, while the loss when choosing $x = t$ is Kt^β. When $K > 1$, we get

that $Kt^\beta > t^\beta \geq t$ for any $t \leq 1$ and $\beta \leq 1$, implying that $x(1/2) = x(1) = 0$, i.e. the dissent is minimal. Thus, the regime cannot do better by choosing $\beta > 1$, and in fact would do strictly worse by doing so, because $\beta > 1$ implies that $S'(0) = 0$, hence no individual would choose $x = 0$. ■

Lemma 18. *Let $R = 0$, $\alpha = 1$, and suppose that $K \in (1/2, 1]$ and all individuals have either $t = 1/2$ or $t = 1$. Then, for any $\mu \in [0, 1]$, the dissent under $\beta \to 0$ is weakly smaller than under any other $\beta \leq 1$, and is such that $x(1/2) = 0$ while $x(1) = 1$.*

Proof. Suppose that the regime chooses $\beta \leq 1$. From Section A.1.1 we know that all individuals have a corner solution, namely either $x = 0$ or $x = t$. The loss when choosing $x = 0$ is t, while the loss when choosing $x = t$ is Kt^β. For individuals with $t = 1$ the value of β has no effect on their choice, but individuals with $t = 1/2$ are more inclined to choose $x = 0$ the lower is β,because $K(1/2)^\beta$ increases in β. Hence, the overall dissent under $\beta \to 0$ is weakly smaller than under any other $\beta \leq 1$. In particular, when $K \in (1/2, 1]$ and $\beta \to 0$ we get that $K(1/2)^\beta \to K > t = 1/2$, hence the moderates ($t = 1/2$) choose $x = 0$, while the extremists ($t = 1$) would choose $x = t = 1$ because for them $K < t = 1$. ■

Proof of Proposition 6

Proof. From Lemmas 15 and 17, we know that, for any $K \notin (1/2, 1]$, the value of μ has no effect on the regime's choice of β (the result holds in a weak sense). If however $K \in (1/2, 1]$, then the regime would choose either $\beta \to 0$ (see Lemma 18), or $\beta > 1$ that is independent of μ (see Lemma 16). However, which of these two will be chosen does depend on μ. In particular, note that under the former choice the moderates stay silent and the extremists speak their minds (Lemma 18), while under the latter individuals who do not speak their minds choose the same stance $\tilde{x}$ (Lemma 16). Hence, if for a given μ' the regime is better off choosing $\beta \to 0$, then for any $\mu > \mu'$ this stays the regime's optimal choice. To see why, note that under $\beta \to 0$, the share of "new" moderates $\mu - \mu'$ decrease their dissent from from $x = 1$ to $x = 0$, which is the maximal possible decrease in dissent. Hence, if $\beta \to 0$ was the optimal choice under μ', it remains the optimal choice. A similar logic implies that if for a given μ'' the regime is better off choosing $\beta > 1$, then for any $\mu < \mu'$ this stays the regime's optimal choice. Thus we get that, overall, the optimal β of the regime weakly decreases in μ. ■

B Other mechanisms and their applicability for the revolutions in the USSR and Egypt

We will briefly outline here a number of standard mechanisms that have been analyzed in the literature on revolutions and mass movements, and discuss the extent to which they can explain the observations in the USSR and Egypt. We wish to emphasize that even if a particular theory is not able to explain the observations we highlight, it does not mean that those forces are not at play more generally.

One of the main problems for protest movements is that in many cases it is individually optimal not to protest – the collective action problem (Olson, 1971; Tullock 1971). Accordingly, the group that is most likely to start a revolution is the one that is most able to overcome the collective-action problem. This could be due to group size, for example. As has been suggested by Esteban and Ray (2001), the individual elasticity of effort of providing the action compared to the publicness of the award to a successful revolution could determine which group size is most conducive for overcoming the collective action problem. While their mechanism may certainly hold in a broader context, it would not provide an explanation for why this group should necessarily hold one ideology or another. Our theory can thus be seen either as orthogonal to theirs or as providing an explanation for why certain groups are more inclined to start protests holding group size fixed. It should also be noted that in the USSR the starting group was much smaller than in Egypt, without there being any obvious difference in the cost elasticity or publicness of the reward for overturning the old regime, suggesting that size was not the only determinant factor.

Another factor in overcoming the collective-action problem is considered by the resource-mobilization theory (see, e.g., Jenkins 1983 for an overview) where the group with the (broadly speaking) best organizational capacity is more likely to be first out protesting. While it is possible that Yeltsin and his closest group were indeed better organized and had more resources than other groups in the USSR, this does not seem plausible in Egypt. In fact, the Muslim Brotherhood were probably better organized (Goldstone, 2011) and could thus coordinate their actions better than the loose grouping that was the first out on Tahrir Square. The organizational capacity of the initiators in Egypt may indeed have improved due to social media (see ElTantawi and Wiest, 2011; Lim, 2012; and Chwe 1999 and Edmond 2013 for formal modeling), but these technologies were of course also available to the Muslim Brotherhood.

Another possibility is that those starting the revolution are not driven by ideology

but are just "opportunist" revolutionary leaders who pick an ideological platform in order to mobilize the masses in their advantage, as modeled by Shadmehr (2015). Such a framework is not applicable to Egypt where the revolution was leaderless, but may be relevant for describing Yeltsin in the USSR. Was Yeltsin an opportunist devoid of personal ideology? Possibly. But a number of subsequent questions then remain. Why did not an opportunist with a real ideological involvement rise up? And if all potential revolutionary leaders are opportunists, why did not the competition among them bring about a leader with an extreme agenda as predicted by Shadmehr (2015, Proposition 4)?

Another question that arises, and relates to all alternative mechanisms discussed here, is why popular policies would trigger a revolution. Most theories emphasize unpopular policies as the ones thought to trigger protest. For instance, political-opportunity theory generally highlights unpopular changes in public policy and increased grievances (Meyer, 1993, 2004). An explanation that pertains to Egypt specifically, but not to the USSR, is that the middle class initiated the Egyptian revolution because the Neo-liberal policies implemented by Mubarak had hurt them disproportionately (Kandil, 2012). The question then is why the working class, many of whom supporters of the Muslim Brotherhood, did not start the revolution, given that they had been hurt by the liberalizing of the economy at least as much. As Armbrust (2011) wrote in Al Jazeera: "The only people for whom Egyptian Neo-liberalism worked 'by the book' were the most vulnerable members of society, and their experience with Neo-liberalism was not a pretty picture."

A related explanation is that the middle class in Egypt had been provided, over a number of years, with higher education but with no jobs to match it (Campante and Chor, 2012; Goldstone, 2011). This could indeed be an explanation for the protests in Egypt. The argument of education further relates to the mechanism of Chen and Suen (2017) that the middle class, with its better education, is more informed about its opportunities than the working class. This indeed seems plausible though probably the leadership of the Muslim Brotherhood would have been at least as informed as the middle class.

One possibility is that the Muslim Brotherhood did not start the revolution because it had the intention to first Islamize Egyptian society before taking power, but that it noticed that the revolution was happening without them, so joined out of fear of losing the window of opportunity. The question then is why the Muslim brotherhood would have such a strategy but not the other groups. One possible answer is given by our model – that extremists cannot speak their minds (so must transform society in more

covert ways) due to the interaction of the ideological cost and the sanctioning structure.

The demographic profile of the population may also play a role for the likelihood of a revolution. This factor, which essentially goes back to McAdam's (1986) argument of demographic availability, may have indeed been potent in Egypt having a large population of youth (Korotayev and Zinkina, 2011). This does not, however, explain why any particular group should be the one starting a revolution, unless this group has a younger age profile than other groups. In fact, this is particularly unlikely to be the explanation for the Egyptian revolution, given that normally the working class and religious groups (who did not start the revolution) tend to have more children than the middle class.

Finally, one possibility is that a starving population, despite having more reasons to protest, has less energy to do so. An increase in economic conditions would then make them more likely to protest. This description does not seem to fit neither the USSR nor Egypt since real hunger was uncommon there (in particular among those who eventually started the revolution). Perhaps more importantly however, empirical evidence (Jia, 2014; Narciso & Severgnini, 2016) seems to speak against this mechanism, as hunger tends to *increase* the likelihood of protests.

There also exist mechanisms that are captured by our model and that may have contributed to the starting of the revolution. There seems to be a consensus that the spark of the revolution in Egypt was the adjacent Tunisian Revolution, which initiated the whole Arab Spring. One possible reason for this contagion in protests is that beliefs about the regime's strength or legitimacy changed after the largely successful protests in Tunisia were observed. This has been modeled formally as a domino-theory of protests between countries by Chen and Suen (2016). In our theory this would be captured, in a reduced form, by a shock to $\bar{K}$ that leads to lower approval or by a direct shock to the approval itself (A). Another possible explanation is that the Tunisian Revolution, and the pictures observed from there, increased the emotions of the Egyptian population, essentially a non-rational-choice argument as has been emphasized by, e.g., van Stekelenburg and Klandermans (2007, ch. 5) and Aminzade (2001). Increased emotions would in our model imply a shift in the relative weight towards the private preferences of the individuals and away from the sanctioning, which is equivalent to a reduction in $\bar{K}$.