From Peer Pressure to Biased Norms: Formation and Collapse*

Moti Michaeli† & Daniel Spiro‡ November 2014

Abstract

This paper studies the emergence of social norms when individuals have heterogeneous private preferences and pressure each other while declaring a stance in public. It characterizes conditions under which a social norm can (and cannot) arise. Further, it shows that peer pressure may lead to a norm which is biased with respect to private preferences in society, yet is endogenously upheld by the population. Strikingly, a biased norm will often be more sustainable than a representative norm, which may explain the bias of various social and religious norms. The model is further applied to political settings, where our analysis of norm stability has implications for revolutionary movements, by predicting which events may initiate a revolution against a political regime and whether the revolution will start with fierce opposition or with mild reform suggestions. The results explain a prevalent yet previously unexplained class of revolutions.

Keywords: Peer pressure, Social norm, Revolution, Protest movement, Alienation, Religion.

JEL: D02, D03, D72, D74, Z10, Z12.

*We wish to thank Martin Dufwenberg, Tore Ellingsen, Joan Esteban, Bård Harstad, Paul Klein, Charles Manski, Arie Michaeli, Kalle Moene, Manuel Oechslin, Andrew Oswald, Paolo Piacquadio, Gerard Roland, Moses Shayo, Sareh Vosooghi, Jörgen Weibull and seminar participants at George Washington University, Hebrew University, Oslo University, Tilburg University, BI and the nordic conference on behavioral economics for valuable comments.

†Department of Economics, European University Institute, Italy. Email: motimich@gmail.com.

‡Corresponding author, Department of Economics, University of Oslo, Norway. daniel.spiro@econ.uio.no, Tel: +47 22855137, Fax: +47 22855035.

1 Introduction

In many social settings, individuals feel pressure to behave in line with their peers. People typically like to have children at the same age as their friends; to drink as much as their peers; and to follow religious customs to the same extent as their co-believers. Minimizing social pressure in these situations is often more complex than just following a social norm. For example, if a Muslim girl has one friend wearing the Burqa, another wearing the Hijab, and a third with no headwear, she will need to trade off conformity between these different friends. Furthermore, as she probably has her own private preference, the existence of peer pressure will force her to trade off the cost of behaving in a way different than her own bliss point with the social loss of deviating from the behaviors of others.

Meanwhile, when behaving in a certain way in public, an individual also indirectly affects others. In modeling terms this means that – when each person takes a stance balancing her private opinion and the peer pressure – we get many stances that in the aggregate shape the social pressure and hence what is considered to be normative.[1] Such modeling lends itself to analyzing the potential emergence of an endogenous norm – a mode of behavior actually followed by many individuals despite their heterogenous preferences and despite them having the option to choose heterogenous stances. This way, the very existence of a norm is not simply assumed but is an equilibrium outcome contingent on people behaving according to it. The previous literature on norms (to be surveyed in the next section) either explicitly assumes a norm exists or implicitly does so by letting the choice space be binary, in which case there must be clustering by construction. By contrast, we let the individual choice set be continuous. This is important as it enables us to show under which circumstances a society can uphold a social norm in equilibrium, and under which it cannot. We furthermore analyze the dynamics leading to the formation and collapse of a norm.[2]

We are particularly interested in analyzing the emergence of a biased norm – i.e., a mode of behavior that is far from the average private opinion yet is followed by

[1] Throughout the paper we treat a declared opinion or stance as synonymous to an action or mode of behavior.

[2] The norm we analyze is *descriptive*, as it depicts a stance chosen by a significant number of individuals. The social psychology literature (see e.g. Cialdini et al, 1991; Cialdini, 2003; Blumenthal et al 2001) distinguishes between descriptive norms (what people do) and prescriptive norms (what people should do). We focus on descriptive norms in the body of the paper but analyze the relation between the two kinds of norms in the appendix.

many. Biased norms are commonplace in social and political life. This has been documented in excessive drinking among college students (for a review see Borsari and Carey, 2001), in attitudes towards alcohol prohibition (Robinson, 1932; Cohen, 2001) and towards racial segregation in the US (O'Gorman, 1975; Fields and Schuman, 1976; Miller and Prentice, 1994), among religious communities (Schank, 1932) and vegetarians (Kitts, 2003), in honor cultures and honor killings (Colson, 1975; Gladwell, 2000; Milgram, 1992; Wilson and Kelling, 1982; Centola et al., 2005), and in norms of violence (Cohen et al., 1996; Vandello and Cohen, 2000).[3]

We divide societies into two classes. One class that can uphold a norm in equilibrium and one class that cannot. The class that *can* uphold a norm can be further divided into two fundamentally different equilibrium societies depending on the underlying preferences. In the first type, which we call an *alienating society*, most (or all) individual statements are identical, creating a norm that few publicly question. Furthermore, if some do question the norm, it will be those who disagree with it the most, openly expressing their very critical private opinions – in that sense being alienated. This means individual non-conformity arises if there is large misalignment between an individual's private opinion and the norm. Thus, a biased norm will be less sustainable than a central norm, as it generates misalignment with many individuals.

In the second type of society that can uphold a norm, those voicing their disagreement with the norm are, somewhat surprisingly, individuals who only slightly disagree with it. Thus, on the surface, one may notice only mild critique of the norm, i.e., a form of internal opposition and debate. But underneath, a larger discontent is concealed, as those who dislike the norm the most choose to fully conform. Moreover, by conforming they unwillingly help to maintain the norm.[4] We call this an *inverting society*, as public conformity and private conformity are inverted. Here, the norm draws its strength from those who privately disagree with it the most as they are

[3]The description of individual behavior in these papers closely resembles our model – an individual, in her will to avoid pressure on herself, indirectly puts pressure on others and thereby takes part in upholding a biased norm.

[4]We implicitly assume that people do not have the option of refraining from declaring a stance. This is standard in the literature (e.g., Kuran, 1989a; Granovetter, 1978; Bernheim, 1994; Manski and Mayshar, 2003; Kuran and Sandholm, 2008; Rubin, 2014), and is directly applicable to situations in which staying silent is either literally impossible (as in the case of choosing headwear) or has a similar peer effect as fully conforming (as in the case of passive obedience). In these situations, the only way of not declaring a stance is to emigrate. In this case the implicit assumption is that emigration is too costly.

the ones who conform. Since a biased norm implies extensive private disagreement, it then follows that a biased norm can survive under *weaker* conditions than a non-biased norm. It will also be dynamically more stable. Hence, the inverting society is particularly suited to uphold biased norms of the kind exemplified earlier. To show that inversion of preferences is not merely a theoretical construct, we provide (in Section 5.1) observations of actual behavior consistent with inversion when it comes to sexual preferences and to religion.[5]

Our analysis has implications also for the sustainment and collapse of political regimes. One standard way of modeling regime stability in the presence of social pressure is to assume that individuals face the binary choice between supporting the existing regime and protesting against it (e.g., Kuran, 1989a; Granovetter, 1978). If the support for the regime decreases substantially the regime collapses, and this is interpreted as a revolution. By interpreting the norm as a regime, we use our framework to further allow individuals to choose the extent of support of the regime – they can completely support it or criticize it to any extent they want on any side of the political spectrum. Our static analysis then shows when a regime (i.e. a cluster of public opinions) can be sustained even in the absence of a group with coherent (private) interests and under which conditions the regime may be biased.[6] It further shows who the regime supporters will be and what private opinions and public statements will characterize the opposition. Our dynamic analysis shows what triggers a revolution, what views those first out to criticize the regime will have, what they will state and which individuals will follow. It further shows when a revolution will start at only one side of the political spectrum and at which point, if any, individuals at the other political extreme will abandon the regime.

We characterize two differnt types of revolutions. In alienating societies, a shift of private sentiments away from the regime will eventually create a pocket of fierce opposition that will spark the revolution. This will gradually induce more moderates

[5]In that section we also provide microempirical support for the parameter assumptions necessary to create an inverting society in our model.

[6]Indeed, as Kuran (1989a, 1989b, 1995) points out, some regimes such as the former Soviet Union remain in power even though they do not represent people's preferences. He argues that this is partly thanks to what seems to be, from the point of view of each individual, a fairly extensive support of the regime by other individuals. This sort of peer pressure was possibly at play also under Hitler's Nazi regime. E.g., Arendt (1964) documents that in the absence of peer pressure, German officers assigned to serve abroad stopped supporting the Nazi regime. She concludes that the "ideal of toughness" concealed a "ruthless desire for conformity" (p.175). See Cohen (2001) for a further discussion and opposing views.

to join. But these moderates will not align their statements with the most fierce opposition but rather speak their own views. Furthermore in this case the revolution will start only on one side of the political spectrum, but at an intermediate stage also those on the other extreme will stop supporting it and instead speak their minds. The last ones to leave the regime will be those on both sides who closely agree with the regime. All in all the revolution will propagate *from the outside towards the inside*. This seems to be a reasonable description of, for instance, the Iranian revolution in 1978-79. This revolution followed a growing misalignment between the Shah and the religious sentiments in society and was initiated by the hardest opponents of the Shah, but then gained mass support by recruiting individuals with more moderate views (Razi, 1987).[7]

The other type of revolution occurs in inverting societies. Here, if the regime's policies become *more aligned* with the private preferences of the population, a revolution may be triggered. Initially, there will be critique from those who nearly agree with the regime on both sides of the political spectrum. These individuals will suggest only mild reforms. This will, however, trigger new and gradually more fundamental suggestions, rejecting the regime. Hence, here the revolution will go *from the inside towards the outside*. This way the inverting society seems particularly interesting to analyze, as it predicts a pattern resembling the sequence of events leading to the collapse of the communist regimes in eastern Europe and possibly also to the recent collapse of Mubarak's regime in Egypt.[8] We treat these in more detail in Section 6.

In the next section we relate the paper to earlier research. In section 3 we present the model and characterize societies that cannot uphold a norm in equilibrium. Sections 4 and 5 analyze the alienating and inverting societies respectively. Section 6 applies the model to political regime formation and revolutionary movements and relates the results to contemporary revolutions and mass protests. Section 7 concludes. The appendix presents some auxiliary results and all formal proofs.

[7]Roughly speaking, revolutions in the alienating society look like those analyzed by Granovetter (1978) and Kuran (1989a), in the sense that the more one opposes the regime, the earlier one is likely to join the revolution against it.

[8]Indeed in Egypt, the historically most extreme opponents of the regime (the Muslim brotherhood) were initially absent from the streets. Furthermore, like our model suggests for the inverting society, the "moderates" in Egypt who did suggest reforms were trying to pull the regime in opposite directions (i.e., some toward more conservatism and others toward more liberalism and openness to the West), implying that the revolution was two-sided initially.

2 Related literature

This section briefly outlines some strands of related literature and how the current paper may contribute to them. Given that social norms, political regime formation and revolutions are vast topics of research, spanning over many disciplines, this description will by no means be exhaustive.

A large part of the previous literature on social norms and conformity to peer pressure is confined to binary stances (e.g., Lindbeck et al., 2003; Brock and Durlauf, 2001; Lopez-Pintado and Watts, 2006; Kuran 1989a; Granovetter, 1978; Angeletos et al., 2007; Acemoglu and Jackson, 2011). Alternatively, when allowing continuous stances, it often takes the norm as exogenous (e.g., Bernheim, 1994).[9] This naturally limits any investigation of endogenously formed norms and their potential bias. Two exceptions are the models by Clark and Oswald (1998) and Michaeli and Spiro (2014). There the location of the norm is determined by the average stance taken by individuals, but this also means that the existence of a norm is assumed rather than derived. To the best of our knowledge the current paper is the first to analyze the endogenous existence of a norm in situations where individuals are heterogenous and can choose heterogenous stances.

When thinking about why social pressure arises, one possibility is that it applies to individuals deviating from a mode of behavior which all agree is appropriate (e.g., being polite or working hard). In this case a prescriptive norm exists exogenously (McAdams, 1997; Cialdini et al., 1991).[10] Another possibility, which is what our paper analyzes, typically concerns situations of ideology, religion or more generally situations where there is a true disagreement about what is right and what is wrong. Here we see no reason why the existence of a norm should be assumed. The papers that are most closely related to ours from this modeling perspective are Manski and Mayshar (2003), where the choice of the number of children of one person depends on the choices of others; Kuran and Sandholm (2008), who analyze the integration speed

[9] In Bernheim (1994) and in Bénabou and Tirole (2006) individuals are punished for what *type* they are preceived to be, which leads to a signling game. By contrast, in our paper and in most of the previously mentioned papers, individuals are punished for their actual actions.

[10] This is the case for example in models of status or work effort (Kandell and Lazear, 1992; Clark and Oswald, 1998; Dufwenberg and Lundholm, 2001). In models of peer pressure like ours, where there are many sources of pressure, a prescriptive norm could, however, also be interpreted as what minimizes social pressure even in the absence of consensus. Then one may analyze a descriptive norm and a prescriptive norm in the same setup. We analyze the relation between these two concepts in our model in more detail in the appendix.

of groups with different preferences; and Acemoglu and Jackson (2014), who analyze the interaction between laws and people's behavior. In these papers, just like in ours, pressure arises between all pairs of individuals in society.[11] All these papers use a quadratic disutility of deviating from one's bliss point, combined with a quadratic pressure when deviating from each other's statement. This is analytically convenient, but it also directly implies that only the average statement in society matters. We use a similar model setup but generalize it so that the quadratic form is contained as a special case. In fact, as we show, no norm can be sustained endogenously under the double quadratic case. So while a quadratic assumption may be of no consequence for their analysis, from the point of view of analyzing the very existence of a norm, it would be very limiting. Furthermore, while quadratic costs of bliss point deviation may provide a reasonable description of preferences in some cases, a number of experiments show that this cost is in fact concave when it comes to ideology and pretence (see Kendall et al., 2013; Gneezy et al., 2013; Gino et al. 2010). The concave case has also been argued to be plausible in ideological and political settings (Osborne, 1995). Similarly, experimental research has found that social pressure is in some societies convex and in some concave (Krupka & Weber, 2013; Hermann et al, 2008; see Michaeli & Spiro, 2014, for further discussion). Given these observations it seems well warranted to take a step towards generalizing the model to include different curvatures than the quadratic. This serves us in pursuing the main goal of our paper – analyzing which societies can, and which cannot, sustain a norm, and when it is likely to be biased.

Our paper also relates to the literature on revolutions and sustainability of political regimes. Following Tanter and Midlarsky (1967), there are two categories of revolutions. Firstly, *coup detats,* performed by elites or a competing party.[12] The second category, which is more related to our paper, is labeled by Tanter and Midlarsky (1967) as *major revolutions,* driven not by a small group of elites but by popular protest and large social movements. Examples of these are the French revolution, the toppling of the Shah in Iran in 1978-79, the collapse of the communist regimes in Eastern Europe and the recent Arab spring. Many such popular protests contain

[11] Also Akerlof (1997) solves a model similar to ours, but restricts the attention to the case where there are only three individuals that are all affected by the statements of each other. In this case it is hard to talk about norms in a formal way.

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

clear ideological or religious motives (Esteban and Ray 2011).[13] For instance, the Iranian revolution (Razi, 1987), the revolutions in Eastern Europe (Lohmann, 1994), the rise of radical Islam (Beck, 2009) and the social movements in Western Europe (Kriesi et al., 1992). Many important insights regarding these environments can be gained by analyzing a simple model where individuals have the binary choice between giving support to the current regime and protesting (see Kuran, 1989a, 1989b, 1995; Granovetter, 1978; Naylor, 1989; and Rubin, 2014). But the binary approach also has some important limitations. For instance, it either assumes or predicts the revolution will be started by those who dislike the regime the most. But many revolutions (to be discussed in Section 6) seem to start by relatively moderate individuals speaking out and our continuous model of revolutionary activity provides an explanation why and when this may be expected. In Section 6 we discuss a number of other novel insights at length and relate them to various real social movements and revolutions.

3 A model of peer pressure and single norm equilibria

We model society as a continuum of individuals, each having a different bliss point $t \in T \subseteq \mathbb{R}$. I.e., some private preference, ideology or opinion, referred to also as the individual's *type*. One can think of t as a position on a political scale. Let $f(t)$ denote a continuous probability density function of types. Each individual publicly declares a stance, visible to everyone else. The publicly declared stance of a type t is her choice variable, denoted by $s(t)$. The inner disutility of an individual declaring some stance s in public, $D(|t - s(t)|)$, increases in the distance between that stance and the individual's type, representing the cognitive dissonance or displeasure felt by her.

In addition, an individual who takes s as a stance feels social pressure $P(s)$. The properties of $P(s)$ are determined endogenously by the model in the following way. When one individual states s and another individual states s', the pressure arising in between them, $p(|s - s'|)$, is increasing in the distance between the stances. Such pressure arises between each pair of individuals. This means that, given a set of stances in society $S = \{s'(\tau) : \tau \sim f(\tau)\}$, the aggregate pressure (P) felt by an

[13] Alternatively, the driving force behind these revolutions may be economic, with the common result that the poorest in society are the ones revolting (see e.g., Stouffer et al., 1949; Merton and Kitt, 1950; Festinger, 1954; Davies 1959; Davies 1962; and more recently for instance Tarrow, 1998; McAdam et al., 2001; Almer et al., 2013).

individual declaring some stance s is given by[14]

$$P(s;S) \equiv E\left[p\left(|s-s'|\right)\right] = \int_{t\in T} p\left(|s-s'(\tau)|\right) f(\tau)\, d\tau. \tag{1}$$

The optimization problem of the individual of type t is about how to minimize the total disutility or loss that arises from the inner disutility and the aggregate social pressure.

$$\min_{s} L(s;t,S) = D\left(|t-s|\right) + P(s;S) \tag{2}$$

All individuals move simultaneously and hence take the stances of others as given. Finding an equilibrium distribution of stances requires solving a fixed point problem, whose solution is a complete mapping from t to $s^*(t)$, where

$$s^*(t) = \arg\min_{s} \left\{P(s;S^*) + D(s;t)\right\} \tag{3}$$

$$\text{s.t. } \left\{S^* : \tau \rightarrow s^*(\tau)\right\}. \tag{4}$$

That is, each individual chooses her stance ($s^*(t)$) optimally given the stances of all others (S^*) such that the chosen stances recreate the ones taken as given by the individual. Being interested in studying the emergence of a norm in society and in the conditions under which this norm may be biased, we first define what we mean by a norm.

Definition 1 *A social norm is a statement $\bar{s}$ made by a non-zero mass of agents. If the social norm is not equal to the average private opinion in society the norm is said to be biased.*

In essence, we require that for an opinion to be called a norm, it should actually be stated by a non negligible number of individuals. In this sense the norm is *real* or, as is denoted in sociology, *descriptive* (Cialdini et al., 1991; Cialdini, 2003; Blumenthal et al., 2001). In Appendix A we discuss the implication of the model for the existence of a *prescriptive norm*, which puts the focus on stances that are approved in society (i.e.,

[14] There are two ways to interpret equation (1). Either s is a statement or action made in public, so that everyone can compare themselves with, implying that $P(s;S)$ is an actual pressure felt when stating s. Or, alternatively, $P(s;S)$ is the expected pressure felt when not knowing whom one is about to interact with following random pairwise matching as suggested by, for instance, Kuran and Sandholm (2008). This formulation is also similar to Esteban and Ray (1994) who use it to measure polarization. However, they are silent as to why individuals choose different stances.

reduce social pressure) yet are not necessarily followed in practice. Being interested in issues of ideology or religion, where there are true differences of opinion with respect to what is the right thing to do, our view is that a real descriptive norm is the most interesting object of analysis.

In order to study the emergence of a single norm, we will confine our analysis to the following type of equilibrium.

Definition 2 *A single norm equilibrium is a solution to the problem in (3) and (4) such that there exists one and only one social norm.*

Note that the continuity of $f(t)$ excludes cases where a norm exists simply because it represents the private opinion of a mass of people. But, to be up-front, the single norm equilibrium is not the only one that may exist in this model, as it may yield more than one norm. But we will confine the analysis to cases where only one norm exists (and to the case where no norm can exist). Wherever applicable, we will perform the analysis for power functions of the form

$$D = |s - t|^{\alpha}, \tag{5}$$

$$p = K\,|s - s'|^{\beta}, \tag{6}$$

where $\alpha > 0$ and $\beta > 0$ represent the curvature of cognitive dissonance and pairwise pressure respectively. K represents the relative weight of the peer pressure, and so captures the extent to which individuals care about social pressure. In our analysis, α, β and K are identical across individuals in a given society.

We start the analysis by characterizing a class of societies in which single norm equilibria cannot exist.

Proposition 1 *If $\beta > 1$ there exists no single norm equilibrium.*

The proof of the proposition appears in the appendix, but the intuition is rather straightforward. For a norm to exist it is required that (some) people will actually state it. But if $\beta > 1$ and if there is clustering of statements at the norm, P will be convex in a neighborhood around the norm and will have a derivative of zero at the norm itself.[15] But with a zero derivative it becomes pointless to state the

[15] Because $p'(0) = 0$ and because otherwise the pressure is not minimized there and so it will not attract a mass of people. There are some subtleties here that are accounted for in the formal proof. The special case of $\alpha = 1$ requires a slightly different intuition but is covered in the proof.

norm exactly, as a small deviation in the direction of one's private opinion reduces the dissonance without increasing the pressure. This means that there cannot be clustering in the first place.

This case, where $\beta > 1$, represents a society where individuals are liberal in how they perceive other's opinions, in the sense that tension (p) arises in between two individuals only when they take distant stances. In a society that consists of such liberal individuals there will never be any reason for two individuals to make the same statement (unless they happen to privately agree). Hence, also at the aggregate level there will not be any one stance that many take. Note also that $\beta > 1$ nests the special case of a double quadratic function as has been analyzed by Manski and Mayshar (2003), Kuran and Sandholm (2008) and Acemoglu and Jackson (2014). Their analyses have different focus then ours, but the previous proposition shows that such a society cannot contain a norm that is upheld by the population. While we find this impossibility result interesting in itself, this paper focuses on norms in equilibrium so we will not delve deeper into the analysis of this class of societies here.

The complementary case is when $\beta \leq 1$, hence p is weakly concave. In this case P may be concave too and hence enable clustering. But note that a concave p does not directly imply a concave P as, depending on the distribution of stances, a concave p may imply also a convex P.[16] However, if sufficiently many follow a norm $\bar{s}$, then that cluster of individuals along with $\beta \leq 1$ will imply a concave P around $\bar{s}$, which is necessary for inducing full conformity by at least some individuals in the first place. This alludes to the existence of a single norm equilibrium in this case.

When $\beta \leq 1$ we find that there exist two qualitatively different types of single norm equilibria. They are treated in great detail in the next two sections. The first type of equilibrium (or society) is one that endogenously induces conformity by those who privately nearly agree with the norm, while alienating those who privately disagree with it strongly. We call this an *alienating* society. It emerges when $\beta < \alpha$ (of course provided that $\beta \leq 1$) and is covered in Section 4. The second type of society is one that endogenously induces conformity by those who privately dislike the norm the most. We call it an *inverting* society. It emerges when $\alpha < \beta \leq 1$ and is covered in Section 5. Together with the class of societies that cannot uphold a single norm equilibrium (described in the previous proposition), the alienating and

[16] For instance, in the appendix (Lemma 5) we show that if all types would speak their minds ($s(t) = t$) and $t \sim U$ then a convex P would arise independently of the curvature of p.

inverting societies span the entire parameter space of α and β. Thus, which one of the alienating or the inverting society will emerge depends on which of α and β is the smallest.[17] To make analytical headway and for brevity we will let the smaller of the two parameters approach zero. I.e., the alienating society will be illustrated by assuming that p is a step function ($\beta \to 0$) and the inverting society will be illustrated assuming that D is a step function ($\alpha \to 0$). The usage of a step function in each case does not drive the results.[18]

We will furthermore assume that the distribution of types is uniform: $t \sim U(-1, 1)$. This of course makes the problem more tractable. But more importantly, it also ensures that a biased norm, following the above definition, does not arise as an artefact of the distribution of types being non-symmetric. We will illustrate and discuss in the appendix (Section B) how our main conclusions translate to other distributions. With a uniform distribution in [-1,1], following (1) and (6), the aggregate pressure function becomes

$$P(s; S) \equiv \frac{1}{2} K \int_{-1}^{1} |s - s'(\tau)|^{\beta} \, d\tau. \tag{7}$$

4 Alienating societies

This section deals with the case where individual pressure (i.e., the pressure arising between two individuals) is concave ($\beta \leq 1$) and, in particular, more concave than D ($\beta < \alpha$). To capture this, suppose p is a step function

$$p(s; s') = \begin{cases} K \text{ if } s \neq s' \\ 0 \text{ if } s = s' \end{cases} \tag{8}$$

while $D = |s - t|^{\alpha}$ for some $\alpha > 0$. A first useful result then follows.

Lemma 1 *Suppose that p is given by (8), D is given by (5) with $\alpha > 0$ and that a single norm $\bar{s}$ exists and is stated by a share x of the population, while the rest speak their minds.*[19] *Define*

$$y \equiv (xK)^{1/\alpha}. \tag{9}$$

[17] To avoid technicalities, we will not analyze in this paper the special cases of $\alpha = \beta$.

[18] We have solved a large part of the general cases analytically and verified the rest numerically.

[19] Throughout the paper, by "speaking ones mind" it is meant that $s(t) = t$.

Then for an individual with private opinion t, the optimal stance is given by

$$s^*(t) = \begin{cases} \bar{s} \text{ if } |t - \bar{s}| \leq y \\ t \text{ otherwise} \end{cases}. \tag{10}$$

This is a partial equilibrium result showing what stance each individual will choose to state given the existence of a norm $\bar{s}$ that is declared by a share x of the population. Here is the intuition. Since p is a step function (and t is continuous), the aggregate social pressure function that results is simply

$$P = \begin{cases} K \text{ if } s \neq \bar{s} \\ (1-x)K \text{ if } s = \bar{s} \end{cases}. \tag{11}$$

The step function is an extreme case that is helpful in capturing the effect of a very concave individual pressure. That is, if the only way to avoid being pressured by someone is to fully agree with her, then the only way to lower aggregate pressure to any meaningful extent is by stating an opinion stated by many. When a single norm exists this could be achieved (only) by stating the norm. Furthermore, since all stated opinions but the norm yield roughly the same pressure when p is very concave, the only effect of the pressure is in determining how unpleasant it feels to state any of these opinions relative to stating the norm. Given such a social pressure function P, the only sensible thing to do for an individual is to either state the norm (thereby lowering pressure) or state her type (thereby not feeling cognitive dissonance). Any other choice will induce some cognitive dissonance while not reducing social pressure. Moreover, two individuals of different types face the same reduction in pressure when stating the norm, but differ in the cognitive dissonance that accompanies such a statement. Thus follows the behavior depicted by the lemma – a type far from the norm will speak her mind while a type close to the norm will declare the norm. y then captures the distance between the norm and a type who is indifferent between these two corner solutions. Overall, this implies that in societies with very concave individual pressure, the ones who dislike the norm the most will be the ones deviating from it in public. In a sense they will be *alienated*.

The previous lemma starts by assuming that individuals divide into two distinctive kinds – those who follow the norm and those who speak their minds – and shows that the same qualitative division is obtained after inducing the individual choices. This hints at the possibility of an equilibrium. However, the actual existence of an

equilibrium hinges on the share of norm followers implied by (10) being equal to the value of x that is assumed in the lemma. In order to establish this relation, the following lemma presents the share of norm followers given the individual optimization in (10).

Lemma 2 *Suppose $s^*(t)$ is according to (10), for a given value of y. Then the share of individuals stating the norm $\bar{s}$ is*

$$x = \begin{cases} y \text{ if } y \leq 1 - |\bar{s}| \\ \frac{y+1-|\bar{s}|}{2} \text{ if } 1 - |\bar{s}| < y < 1 + |\bar{s}| \\ 1 \text{ if } y \geq 1 + |\bar{s}| \end{cases}. \tag{12}$$

Furthermore, x is increasing in y and decreasing in $|\bar{s}|$.

This lemma presents the share of the population (x) that will choose to declare the norm as a function of y (the distance between the norm and the indifferent type). It builds on the previous result that those close to the norm will fully conform while those far from it will speak their minds. This directly implies that the further from the norm the indifferent type is, the greater is the number of individuals conforming to the norm. The use of a uniform distribution at $[-1, 1]$ implies that when $\bar{s} = 0$ we automatically get that $x = y$, but when $\bar{s} \neq 0$ the mapping from y to x is not one-to-one for every y, as expressed in (12).[20]

A static equilibrium of the model is essentially a fixed point defined by a triplet $(x, y, \bar{s})$ that satisfy Lemma 1 and Lemma 2 simultaneously. The conditions for the existence of such an equilibrium are presented in the following proposition.

Proposition 2 *Suppose that individual pressure is according to (8) and D is given by (5) with $\alpha > 0$. Then:*

1. *For each value of $\bar{s} \in [-1, 1]$ there exists a single norm equilibrium with a norm $\bar{s}$ if and only if K is sufficiently large.*

[20] If the norm is biased, say, to the left ($\bar{s} < 0$), then if y is large enough (in particular, larger than $1 - |\bar{s}|$, the distance from the norm to the left edge of the type distribution), *all* types to the left of the norm declare the norm. Thus, as we increase y further, the only new types declaring the norm will be on the right side of it. Finally, when y is so large that it exceeds $1 + |\bar{s}|$ (which is the distance from the norm to the most extreme type in society) then everyone conforms to the norm, implying that $x = 1$.

2. *Denote the infimum value of K that supports a single norm equilibrium by $K_{\min}(|\bar{s}|)$. Then $K_{\min}(|\bar{s}|)$ is weakly increasing in $|\bar{s}|$.*

This proposition expresses three main results, which hold also beyond the step function case. Firstly, that under sufficiently concave individual pressure there exist single norm equilibria whenever individuals care sufficiently about social pressure – K has to be greater than $K_{\min}(|\bar{s}|)$.[21] Secondly, that in these equilibria the norm may be biased. Thirdly, that the more biased the norm is, the larger is the K needed to sustain it in equilibrium. This last result is a key result. It essentially says that in order to uphold a biased norm, individuals in society need to care about social pressure more than is needed in order to uphold a more central norm. The intuition for this result is that the strength of the norm depends on the number of followers, where potential followers are types with opinions close to the norm. Therefore, when the norm is biased there are more private opinions further away from the norm and hence more potential deviators. To sustain the norm this has to be compensated for by a heavier weight of pressure. This can also be seen in Lemma 2, which states that given y the share of norm followers falls with biasness of the norm.

Figure 1 depicts this equilibrium. The two graphs on the left show the case of a central norm, where the distribution of stances is shown in the upper left schedule and the mapping of types to stances in equilibrium is shown on the lower left. In this particular case all individuals conform fully to the norm. The right graphs show the case of a biased norm. Here a group of extreme objectors express their heterogenous private opinions

The previous results imply that there can be multiple equilibria in the sense that the norm can be located at more than one place. But these equilibria are not different in kind – each of them contains a norm, where types far from it speak their minds while types close to it (sometimes all) conform (we will refer to such distribution of stances by the label *alienation*). The equilibria differ only in the location of the norm and in the share of the population following it.[22]

[21] In the case of $\alpha > 1$ we have $K > K_{\min} = 0$. That the lower bound is zero is an consequence of assuming p is a step function. If one were to assume a concave p (but not a step function), $K_{\min}$ would be greater than zero also when $\alpha > 1$. The rest of the results presented are not specific to the step function assumption but hold more generally when $\beta < \alpha$ and $\beta \leq 1$.

[22] More precisely, Lemma 1 and Proposition 2 say that the form of the distribution of stances in a single norm equilibrium is unique in the sense that a single norm equilibrium is established if and only if the distribution of stances displays a cutoff within which all conform and beyond which all speak their minds.

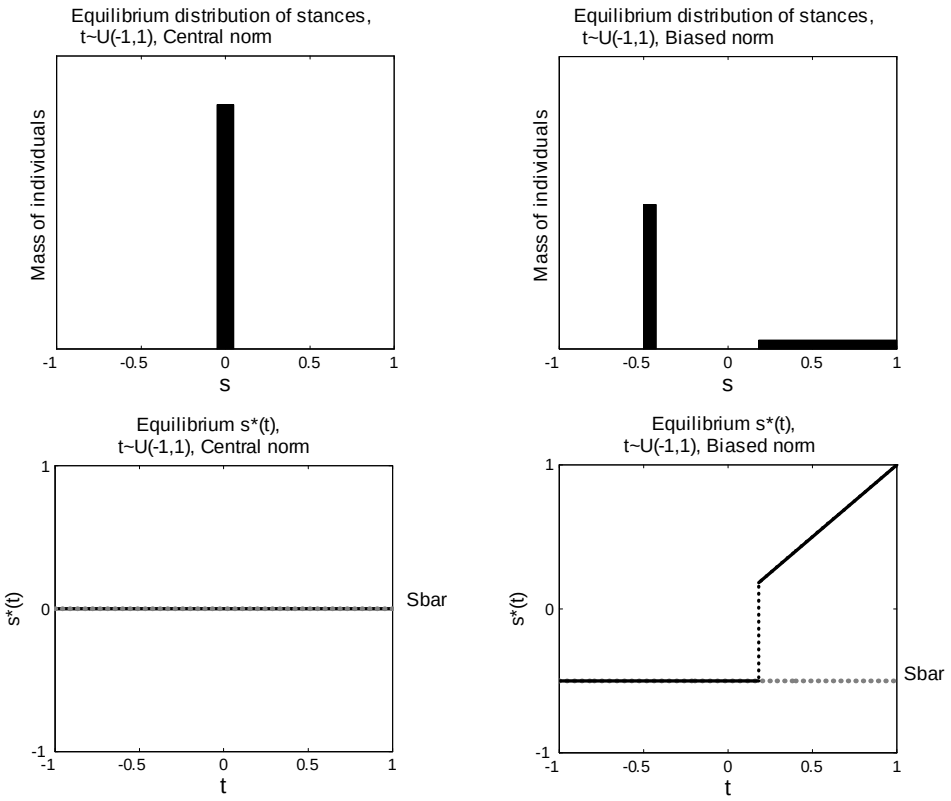

Figure 1: The left graphs show the distribution of stances (top) and $s^*(t)$ (bottom) in equilibrium with a central norm ($\bar{s} = 0$). The right graphs show the distribution of stances and $s^*(t)$ in equilibrium with a biased norm ($\bar{s} = -0.5$). In all figures $\beta = 0.01$, $\alpha = 0.9$ and $K = 1.2$.

It is important to analyze whether these equilibria are merely a possibility or whether they are also stable in a dynamic sense. For this purpose we will now add a dynamic structure to the model. It can be interpreted either as individuals adjusting their statements when observing what others have stated, or as an overlapping generations model, where the stances of the older generation (the parents) create pressure on the younger generation (the kids) when choosing their own stances, and this generation puts pressure on their kids and so on.[23] Let i indicate the period of the dynamic process (representing a period or a generation). Then an individual of

[23]Implicitly we assume here that the distribution of *types* is stationary between generations. For short to medium run analysis (say, limited to at most a few decades), a fixed distribution of types seems not too extreme an assumption. In particular, when thinking about revolutions, as Kuran (1989a) and Granovetter (1978) do. Furthermore, assuming private preferences are not affected by the norm essentially makes it harder to sustain a norm. At any rate, in Appendix B we discuss the extent to which our results should hold under alternative dynamic assumptions too.

type t in period i solves the following problem.

$$\min_{s_i} L\left(s_i; t, S_{i-1}\right) = D\left(|t - s_i|\right) + P\left(s_i; S_{i-1}\right) \text{ where} \tag{13}$$

$$P\left(s_i; S_{i-1}\right) \equiv \frac{1}{2}\int_{-1}^{1} p\left(\left|s_i - s^*_{i-1}(\tau)\right|\right) d\tau.$$

Clearly, any equilibrium found in the dynamic problem will also be an equilibrium in the static problem. But the converse is not necessarily true. A static equilibrium could be practically non-attainable in a dynamic sense. So the dynamic problem will help us rule out equilibria that have no gravity.[24]

Proposition 3 *Consider the dynamic model in (13) with p being a step function as in (8) and D as given in (5) with $\alpha > 0$. Then:*

1. *There exists a stable steady state with a single norm $\bar{s} \in [-1, 1]$ if and only if $K > K_{\min}\left(|\bar{s}|\right)$, where a share $x_{ss}\left(|\bar{s}|\right) > 0$ of the population declare the norm.*

2. *$x_{ss}\left(|\bar{s}|\right)$ is weakly decreasing in $|\bar{s}|$.*

3. *Consider a norm $\bar{s}$ and suppose $K > K_{\min}\left(|\bar{s}|\right)$. Let x_i denote the share of norm followers in period i. Then there exists a value $x_{conv}\left(|\bar{s}|\right)$ such that if $x_i > x_{conv}\left(|\bar{s}|\right)$, there is convergence to a stable single norm steady state with $x_{ss}\left(|\bar{s}|\right) > x_{conv}\left(|\bar{s}|\right)$. Otherwise, if $0 \leq x_i \leq x_{conv}\left(|\bar{s}|\right)$, there is convergence to a stable steady state where each type speaks her mind ($x_{ss}\left(|\bar{s}|\right) = 0$).*

4. *$x_{conv}\left(|\bar{s}|\right)$ is increasing in $|\bar{s}|$ and decreasing in K.*

The proposition highlights that if $K > K_{\min}\left(|\bar{s}|\right)$ and sufficiently many conform to a norm in some period i, then society will converge to a stable steady state where this same norm is upheld endogenously. However, this requires a minimum amount of conformity ($x_i > x_{conv}$) at the onset. If this initial condition is not satisfied, then in each consecutive period more and more people will speak their minds, until all do so and society reaches a state of complete *pluralism*. If K is below $K_{\min}$ to begin

[24] For brevity, in the proposition we will treat the unstable steady states (x_{uss}) as ones where if $x_i = x_{uss}$ then $x_{i+1} < x_{uss}$.

with, then the initial norm cannot be sustained at all, no matter how many declare it initially.[25]

With respect to the properties of dynamic convergence, Lemma 1 shows that alienation is a distribution of stances that recreates itself. That is, if there is a cutoff distance from the norm, beyond which types speak their minds and within which they follow the norm, then there will exist a cutoff also in the next period. This implies that, for a given $\bar{s}$, the full dynamics of the model can be derived by analyzing a function $x_{i+1} = f(x_i)$. This function is the main building block for proving Proposition 3. Here we demonstrate it for the case $\alpha < 1$ in Figure 2. It depicts a phase diagram with x_i on the horizontal axis and x_{i+1} on the vertical axis. The 45 degree diagonal depicts the steady state values where $x_{i+1} = x_i$. As can be seen in the figure, $f(0) = 0$, and then $f(x_i)$ starts below the 45 degree line, but afterwards it increases and crosses the 45 degree line and stays above it (if and only if $K > K_{\min}(|\bar{s}|)$). Hence, $x = 1$ and $x = 0$ are stable steady states, while there is an interior non-stable steady state in-between them. The value of x in this inner state also forms the boundary between the zone of convergence to a single norm stable steady state and the zone of divergence toward a state of pluralism. I.e., this is x_{conv} of the proposition. The figure also highlights that the steady state in which a norm exists ($x = 1$) is stable not only with respect to small perturbations – there is convergence to it from a rather broad range of initial conditions (of course depending on the value of K). In the specific example depicted in the figure, the stable single norm steady state is degenerate, in the sense that everyone in society adheres to the norm ($x = 1$), but more generally there can be stable steady states exhibiting some alienation.[26]

Apart from convergence, the proposition also highlights the effect of the bias of the norm. Parts (1) and (3) of the proposition imply that a biased norm can persist also in a dynamic setting. This means that societies may be history dependent in the following sense. Suppose a group of individuals at some point state the same opinion.

[25] When K equals $K_{\min}$ then the single norm equilibrium is stable only with respect to deviations in which too many initially follow the norm (i.e., it is stable if $x_{ss} < x_i \leq 1$).

[26] Another case which was not depicted in the figure is the one where $\alpha \geq 1$. Then $f(x_i)$ goes immediatelly above the 45% line and potentially crosses it from above implying there is convergence to a single norm steady state for any $x_i > 0$. This very broad range of convergence when $\alpha \geq 1$ is however specific to the step function case. For $\beta > 0$ and $\alpha \geq 1$ the $f(x_i)$ function starts *below* the 45% line and crosses it from below implying the same dynamics as those described in the main text for $\alpha < 1$.

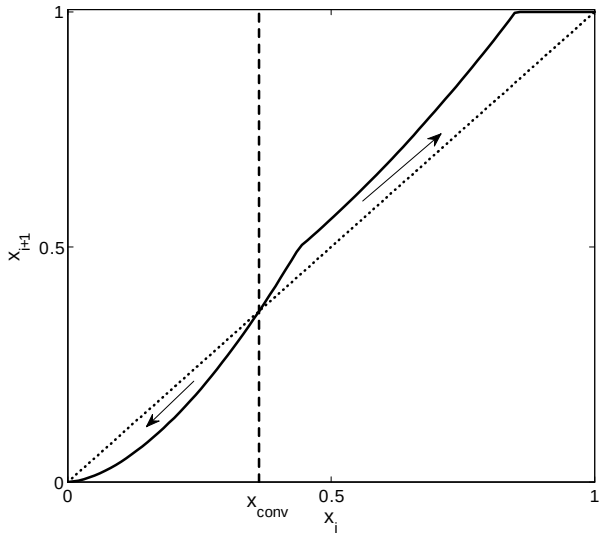

Figure 2: A phase diagram showing convergence to a single norm steady state with $\bar{s} = -0.5$ for p being a step function, $\alpha = 0.6$ and $K = 1.5$. The dotted line depicts the diagonal where $x_{i+1} = x_i$, the solid line depicts the intertemporal dynamics $x_{i+1} = f(x_i)$. The vertical line depicts x_{conv}, i.e., the boundary between the zone of convergence to a single norm equilibrium ($x = 1$) and to "pluralism" ($x = 0$).

Then, provided that they are sufficiently many ($x_i > x_{conv}(|\bar{s}|)$) or powerful, this opinion may be established as a norm and may persist also after those individuals are gone, even if it does not represent the average private opinion in society. Note also that if that initial group is only slightly larger than x_{conv}, the norm will gain more followers over time, thus becoming stronger. The fourth part of the proposition states that the minimum amount of conformity necessary for the norm to be sustainable in the long run is decreasing in the weight of the pressure and increasing in the bias of the norm. This can be demonstrated using Figure 2. By increasing K, the function $f(x_i)$ tilts upwards, which implies that x_{conv} decreases and so the zone of convergence increases. However, by increasing $|\bar{s}|$, the function $f(x_i)$ tilts downwards, implying a smaller zone of convergence. Hence, increasing K and increasing $|\bar{s}|$ works in opposite directions. This means that, while a biased norm *can* exist, the more biased it is, the less magnetic it is, unless it is compensated for by a larger K. Hence, biased norms are less sustainable than central norms in two ways. Firstly, they require people to care more about social pressure ($K_{\min}$ is higher). Secondly, they require more conformity in the first period (x_{conv} is higher). Part 2 of the proposition also states that public

cohesion in society is falling with biasness.

5 Inverting societies

This section deals with the case where the individual pressure p is concave, but less so than the dissonance function D ($\alpha < \beta \leq 1$). To capture this, suppose that D is a step function

$$D(s;t) = \begin{cases} 1 \text{ if } s \neq t \\ 0 \text{ if } s = t \end{cases} \tag{14}$$

while $p = K|s - s'|^{\beta}$ for some $\beta \leq 1$. The following partial equilibrium result describes what stances individuals choose to state given a certain form of social pressure, which will be justified later on.

Lemma 3 *Suppose $P(s)$ is monotonically increasing in the distance from $\bar{s}$, and D is according to (14). Then on each side of the norm there exists a cutoff value such that types closer than the cutoff speak their minds and types further away than the cutoff state $s^*(t) = \bar{s}$.*

This lemma presents the general pattern of individual choices in a society in which social pressure (P) is increasing with the distance from a certain stance $\bar{s}$ and individuals are very perfectionist. Essentially, the lemma says that types close to $\bar{s}$ will speak their minds while types further away will state $s^*(t) = \bar{s}$, thus fully conforming to a unique norm. The intuition for this is rather straightforward. When D is a step function, an individual will either speak her mind, or, once she deviates from her private opinion, say whatever lowers social pressure the most. This is since she does not distinguish between statements that are not exactly her private opinion. The question then is which individuals will be the full conformers and which individuals will speak their minds. When social pressure is increasing with the distance from the norm (while the dissonance of deviation from one's bliss point is independent of type), types far from the norm will find it the hardest to speak their minds. Hence, there will be a unique cutoff such that types further away from the norm than the cutoff point will follow the norm, while types closer will speak their minds. On the aggregate level this can be interpreted as an *inversion of preferences,* whereby those who despise the norm the most are the ones declaring it in public. Meanwhile, those who nearly agree with the norm speak their minds openly, thus criticizing it mildly.[27]

[27]Note that this result is not particular to D being a step function. Roughly speaking, if D is concave, it suffices that it is very concave (small α) and that the aggregate pressure P is concave

Now, the previous lemma was a form of partial equilibrium since it assumed that P is an increasing function with a unique minimum point $\bar{s}$. The question then is whether the individual choices depicted in Lemma 3 induce such properties of P. In the upcoming analysis we will use y (with some abuse of notation) to denote the distance between the norm and the type who is indifferent between speaking her mind and stating the norm.

Lemma 4 *Suppose $\beta \leq 1$ and there exist some $\bar{s} \in [-1,1]$ and $y \in [0, 1+|\bar{s}|]$ such that all types with $|t-\bar{s}| \leq y$ choose $s^*(t) = t$ while the rest choose $s^*(t) = \bar{s}$. Then there exists a value $y_{\max}(\bar{s}) \geq 1$ such that $P(s)$ is monotonically increasing in $|s-\bar{s}|$ if and only if $y \leq y_{\max}(\bar{s})$.*

While the previous lemma described what individuals state given social pressure, this lemma states the properties of social pressure given what individuals state. The bottom line of the lemma is that if types far from the norm follow the norm and those close to the norm speak their minds, then P will be strictly increasing in the distance from the norm as long as there are sufficiently many norm followers. This is the same as requiring that the most deviant opinion expressed in society (at distance y from the norm) is not too deviant. $y_{\max}(\bar{s})$ then measures how critical the most critical opinion can be while still ensuring that P is everywhere increasing in the distance from $\bar{s}$.[28]

Put together, Lemmas 3 and 4 allude to the existence of an equilibrium, since the first says that inversion of preferences will arise if P is increasing in the distance from $\bar{s}$ and the second roughly says that given inversion, P will be increasing in the distance from $\bar{s}$. The conditions for the existence of such an equilibrium are presented in the following proposition.

Proposition 4 *Suppose D is according to (14) and p is according to (6) with $\beta \leq 1$. Then:*

1. *For each value of $\bar{s} \in [-1,1]$ there exists a lower bound on K, denoted by $K_{\min}(|\bar{s}|)$, such that a single norm equilibrium with a norm $\bar{s}$ exists if and only if $K \geq K_{\min}(|\bar{s}|)$.*

close to $\bar{s}$ and increasing throughout. For a result along these lines see Michaeli and Spiro (2014).

[28] For a norm in the center of the type distribution it does not matter how many follow it, as it will be the global min point of pressure anyway (note that $y_{\max}(\bar{s}) \geq 1$). But if the norm is biased and only few follow it, the min point of pressure may be located elsewhere. This sets a bound on the maximum amount of deviation, which is captured by $y_{\max}(\bar{s})$.

2. $K_{\min}(|\bar{s}|)$ *is weakly decreasing in* $|\bar{s}|$.

While the full proof is in the appendix, we will now partly explain it by illustrating some properties of the equilibrium. The potential existence of a single norm equilibrium was explained earlier. The proposition confirms the actual existence of such equilibria whenever individuals care sufficiently about social pressure – K has to be greater than some $K_{\min}(|\bar{s}|)$.[29] The reason for the requirement of a sufficiently large K is that there need to be enough individuals who fully conform in order to make the social norm strong enough to actually be a point of attraction. However, unlike in the alienating society (see Proposition 2), here the pattern of individual choice is that of *inversion of preferences*. Here those who dislike the norm the most (i.e., those furthest from it) declare it and hence are the ones upholding the norm.[30] In the appendix (Lemma 17) we show that inversion is the *only* pattern of individual choice consistent with a single norm equilibrium.

The second part of the proposition implies that a biased norm not only may exist, but also requires weaker conditions for existence than a norm that is more centrally located. To understand why this is the case, recall that a very concave D implies that types far from the norm conform while those close to it state their private opinion. This creates a distribution of types as depicted in the upper left graph of Figure 3. Suppose now that we slightly move the norm towards the left edge. The conformity of types at the edges of the type distribution then implies that the "distribution package" will move together with the norm without changing appearance – those beyond $\bar{s} \pm y$ will fully conform, while those within this range will speak their minds. This shows that biased norms may exist. Now, if we continue moving $\bar{s}$ leftward, at some point the type $t = \bar{s} - y$ will equal -1. When moving $\bar{s}$ beyond this point, the left wing of the uniform part will be truncated (as in the upper right graph of Figure

[29] Note that this value is not necessarily equal to the $K_{\min}(|\bar{s}|)$ of Proposition 2.

[30] The full conformity of these people holds true also for other concave D functions (i.e., not just the step function). The concavity of D assures that types with opinions far from the norm do not distinguish much between fully conforming and stating other opinions that are almost as far from their bliss points as the norm is. In the more general case ($0 < \alpha < \beta$) this pattern of behavior hinges on the aggregate pressure P being not only increasing in the distance from the norm, but also concave around the norm, which results from the existence of a mass of norm supporters who impose concave individual pressure. This is important and non-trivial. Important, since if P is not concave, there is no point for anyone to fully conform. It is non-trivial since the group of *non-conformers* impose together a *convex* aggregate pressure. But, as can be seen by differentiating equation (25) with respect to s and letting $s \to \bar{s}$, P is still concave close to the norm, since there the conformers have a larger effect on pressure, and the contribution of this group is concave.

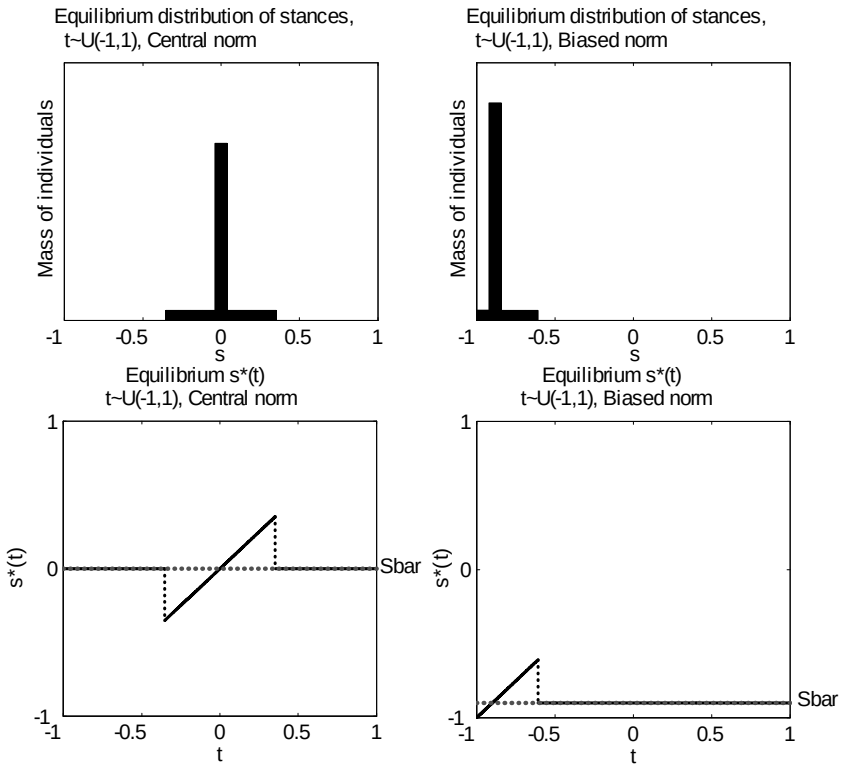

Figure 3: The left graphs show the distribution of stances (top) and $s^*(t)$ in equilibrium (bottom) with a central norm ($\bar{s} = 0$). The right graphs show the distribution of stances and $s^*(t)$ in equilibrium with a biased norm ($\bar{s} = 0.9$). In all figures $\beta = 0.6$, $\alpha = 0.1$ and $K = 1.6$.

3), thus changing the shape of the stance distribution and potentially also affecting the indifference of the type $t = \bar{s} + y$. As will be explained shortly, when the left wing is truncated, y (measuring the size of the right wing) becomes smaller, implying even more conformity in society. Consequently, a lower K is needed in order to sustain the norm in equilibrium. All in all, bias thereby compensates for weakness of social pressure, making biased norms more sustainable than central norms.

For the upcoming dynamic results it is helpful to understand why truncation of the left wing of those speaking their minds induces more conformity from the right. First note that high social pressure in itself is not sufficient to induce a person to deviate from her bliss point. Rather, there needs to be some other stance that lowers pressure substantially enough to make it worthwhile to endure the cost of pretence.

With the fixed cost of pretence (equation 14), we get that conformity is induced when

$$P(t) - P(\bar{s}) \geq 1.$$

Now, the effect of truncation of the left wing of the uniform part resembles that of induced conformity by people on the left side of the norm (where non-conforming types cease to exist due to the truncation). Therefore, suppose indeed that for some reason, a group of types from the left side of the norm decide to follow the norm. This has two opposing effects on the pressure on people on the right side of the norm. On the one hand, it decreases $P(t)$ for every $t > \bar{s}$, so that speaking one's mind is easier, as the statements of the previous leftists are now closer to the right. This has an effect of disincentivizing rightists to conform. But on the other hand, the second effect is that $P(\bar{s})$ decreases too, since there are more individuals stating the norm. When p is concave, this latter effect is stronger (because the concavity of the individual pressure function implies that the reduction in pressure is more substantial at $\bar{s}$ than at any point to the right). Hence, the new indifferent type will be closer to the norm. An interpretation of this would be that conformity of leftists helps conform rightists.

It is important to analyze whether the previous equilibrium with a single norm is merely a theoretical possibility or whether it is dynamically stable. For this purpose we will now add the same dynamic structure to the model as we did in the previous section (see equation 13). Before stating the analytical result, it may be worthwhile to revisit Lemmas 3 and 4. Lemma 3 implies that, if in a certain period the social pressure increases in the distance from its minimum point, then in the next period there will be inversion of preferences, whereby a norm is created at that minimum point. This recreates a pressure that increases in the distance from the norm (Lemma 4), which will again imply inversion by Lemma 3. Hence, inversion with a single norm is a situation that will tend to recreate itself dynamically. The question then is whether this process will settle on a stable steady state where a norm still exists.

Proposition 5 *Consider the dynamic model in (13) with D being a step function as in (14) and p as given in (6) with $\beta \leq 1$. Then:*

1. *There exists a stable steady state with a single norm $\bar{s} \in [-1, 1]$ if and only if $K > K_{\min}(|\bar{s}|)$, where a share $x_{ss}(|\bar{s}|) \in]0, 1[$ of the population declare the norm.*

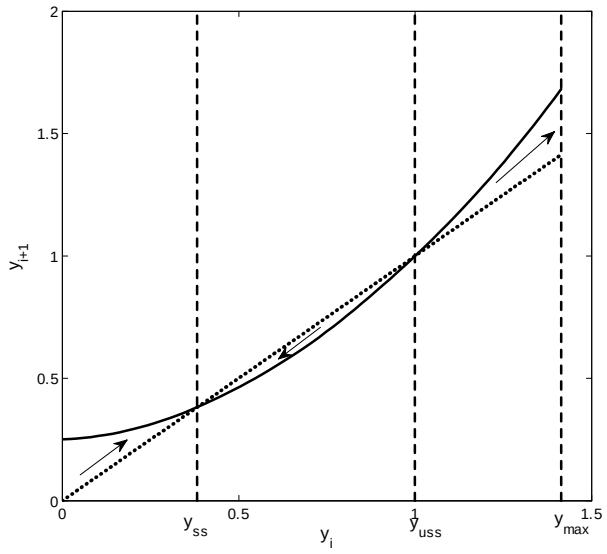

Figure 4: A phase diagram showing convergence to a stable single norm equilibrium when $\bar{s} = -1$, for D being a step function, $\beta = 0.5$ and $K = 2$. The dotted line depicts the diagonal where $y_{i+1} = y_i$, the solid line depicts the intertemporal dynamics $y_{i+1} = f(y_i)$. The vertical lines depict the upper bounds for convergence, y_{uss} and $y_{\max}$ (with y_{uss} being the binding one in the case depicted here). The phase diagram is not defined for $y_i > y_{\max}$.

2. *$x_{ss}(|\bar{s}|)$ is increasing in $|\bar{s}|$.*

3. *Consider a norm $\bar{s}$ and suppose $K > K_{\min}(|\bar{s}|)$. Let y_i denote a cutoff value in period i, such that all types with $t \in [\bar{s} - y_i, \bar{s} + y_i]$ speak their minds while the rest follow the norm. Then there exists a value $y_{conv}(|\bar{s}|)$, such that there is convergence to a stable steady state with a single norm $\bar{s}$ if $y_i < y_{conv}(|\bar{s}|)$.*

4. *$y_{conv}(|\bar{s}|)$ is increasing in $|\bar{s}|$.*

As discussed earlier, the pattern of the dynamic process is such that inversion of preferences in period i recreates inversion in period $i+1$ with a new cutoff value. This implies that the dynamic process can be described solely by the temporal cutoff y_i (the distance between the norm and the person furthest away from it who speaks her mind in period i). Figure 4 shows a phase diagram that depicts y_{i+1} (vertical axis) as a function of y_i (horizontal axis). As can be seen from the figure, there is a stable steady state with a norm when $y_i = y_{ss}$. The existence of such a steady state

for a given $|\bar{s}|$ hinges on K being strictly greater than $K_{\min}(|\bar{s}|)$, as defined in the static Proposition 4.[31] It may be interesting to note that the steady state is never degenerate – there is always a share of the population (those close to the norm) who speak their minds. In the proof of the proposition we show that an increased $|\bar{s}|$ pushes the function y_{i+1} downward, which implies that y_{ss} decreases with biasness, so that the most critical opinion in the steady state becomes less critical. This has the further consequence that the share of the population conforming increases with biasness (part 2 of the proposition). Together, these two observations imply that the cohesiveness of stated opinions in the stable steady state increases with the bias of the norm.[32]

If $y_i < y_{ss}$, society will converge to this stable steady state. Furthermore, there may be another, unstable, steady state at y_{uss}, which marks the border between the convergence zones. Now, the existence of y_{uss} hinges on $f(y_i)$ intersecting the 45 degree line twice to the left of $y_{\max}$. Beyond $y_{\max}$, P is non-monotonic and hence the phase diagram is not applicable there. If there exists such $y_{uss} < y_{\max}$, as depicted in the diagram, then $y_i < y_{uss}$ is a necessary and sufficient condition for convergence to the stable steady state y_{ss}. However, if instead y_{ss} is the unique point of intersection, there is convergence to y_{ss} starting from any $y_i < y_{\max}$. Hence, sufficient conditions for convergence is that $y_i < y_{\max}$ and that $y_i < y_{uss}$ whenever it exists.[33] The last point of the proposition says that the sufficient condition for convergence, $y_{conv} \equiv \min\{y_{uss}, y_{\max}\}$, increases with biasness. This is so because an increased $|\bar{s}|$ tilts the function y_{i+1} downwards, which implies an increase in y_{uss}, and because $y_{\max}$ also increases with biasness. What happens when starting beyond y_{conv}? This is harder to say analytically since then pure inversion may not be maintained. But an extensive set of of simulations of the model for different combinations of $\bar{s}$, K and β constantly shows the same result: there is in practice a maximum value of y_0 within which there

[31] In the case of $K = K_{\min}$ there is convergence to the steady state only from $y_i < y_{ss}$.

[32] This is another manifestation of the contribution of biasness to the sustainability of single norm equilibria. The intuition here is the same as for why biased norms enable a lower $K_{\min}$. Once there are fewer individuals on one side (due to biasness truncating the wing of critique on that side), this makes more individuals on the other side conform. Conformity thus increases from both sides of the norm, and the most critical opinions expressed in public become less critical among leftists and rightists alike.

[33] There can be convergence also when starting from $y_i > y_{\max}$ but showing the precise necessary and sufficient conditions is substantially harder. This is since, beyond $y_{\max}$, the potential convergence will not display pure inversion in every period – there may be several disjoint sets of individuals speaking their minds.

is convergence to a single norm steady state with inversion, and beyond which society converges to pluralism, where each individual speaks her mind. Furthermore, these simulations suggest that this cutoff of convergence is increasing in biasness.

y_{conv} may be interpreted as describing a maximum level of initial public critique. If initially a norm exists and the most critical person is less critical than y_{conv}, then the norm will stay stable over time. This implies that if a group of individuals, possibly a long time ago, declared together one opinion, then this opinion could become an endogenous norm, upheld by those who despise it the most. This holds true even if this group was rather small (i.e., of size $x_0 < x_{ss}$, with $y_0 \in (y_{ss}, y_{conv})$), in which case the norm becomes stronger over time. Hence, we can start with a norm that is to a non-trivial degree weaker than in steady state and still converge back to a steady state with a single norm.[34] What point (4) of the proposition (in combination with our simulations) suggests, is that the most critical opinion in the first period can be more critical the more biased the norm is.[35]

All in all, we get in this section a very different result compared to that of the previous section. Here, when cognitive dissonance is very concave, biased norms not only exist and are dynamically stable, but they are also more sustainable than central norms. This is since, compared to central norms, they require lower social pressure ($K_{\min}$ is lower), imply more cohesion (x_{ss} is higher) and maintain their attraction in the presence of harsher initial critique. Another important difference compared to alienating is that now it is those who nearly agree with the norm who speak their minds, implying that the critique expressed publicly in society will be rather mild.

5.1 Observations of inversion

In this subsection we will briefly present some empirical and casual observations of inversion of preferences taking place. Before doing so, let us just note that recent experimental research implies it is plausible that the combinations of preference curvatures that lead to inversion in our model indeed exist in some cases.[36] Finding

[34] Of course, if this initial group was large to begin with and consisted of the whole society, then the norm will also persist, but will become weaker over time, as some people will deviate from full conformity and speak their minds openly.

[35] For some parameter combinations, there may exist two stable steady states with an unstable steady state in between. In this case, starting with a rather high y_i implies convergence to the higher of the stable steady states while starting with a low y_i implies convergence to the lower of the stable steady states. The statements in the proposition are expressed taking this into account.

[36] First, in order to create inversion of preferences, the disutility of deviating from one's blisspoint must be concave. Recent experimental research provides support for this assumption, showing

real-life examples of inversion of preferences requires knowing the private opinions of people, which are usually unknown. Yet there exist some real cases where private opinions are obtainable, and that are, at least observationally, consistent with our description of inversion of preferences.[37]

One example where inversion may be observed is when considering sexual orientation. To help fix ideas, think of a society where homophobia is normative (in the sense of a lower pressure on declared homophobia vis-à-vis other stances). One can think of three prototypical types living in this society, with a private preference to be either homosexual, or a non-homophobic heterosexual, or a homophobic heterosexual. Now, suppose homophobes state their private preferences in public; (at least some) non-homophobic heterosexuals also present their private preferences in public; and (at least some) homosexuals appear in public as homophobic heterosexuals while never publicly appearing as non-homophobic heterosexuals. This would constitute an inversion of preferences, as people far from the homophobic norm (i.e., homosexuals) state it, while those close to the norm (non-homophobic heterosexuals) speak their minds. Data in support of this behavioral pattern was indeed reported. First, note that many (if not most) societies hold negative attitudes towards homosexuality (Pew research center, 2007; Savin-Williams & Ream, 2003). It is also well documented that not all homosexuals disclose their true sexual preferences (D'Augelli, 2006). What Adams et al. (1996) showed in addition is perfectly in line with inversion of preferences. In their experiment, participants got to state their preferences in a survey (i.e., their stances) and then their homosexual tendencies were measured physically (i.e., their types). The results showed that all those physically measured to be homosexual were either self-reported homosexuals or self-reported heterosexual homophobes,

that individuals behave as if they have concave disutility from bliss point deviations in two relevant situations – when considering ideological stances that differ from their own (Kendall et al., 2013) and when considering cheating (Gneezy et al., 2013; Gino et al. 2010). Second, the individual pressure needs to be concave too, but less so than the disutility from bliss point deviations. The existence of societies characterized by such pressure is indeed plausible in light of the large intercultural differences in peer pressure documented in Herrmann et al. (2008). In particular, they report how individuals punish others who contribute a different amount than themselves in the public good game and find a vast heterogeneity in the curvature of peer pressure, ranging from a very convex pattern in Melbourne, to slightly concave in Riyadh and Muscat and to a very concave one in Athens (see Figure 1 in Herrmann et al. 2008).

[37] When considering inversion in real-life scenarios, it should be interpreted in a loose sense. I.e., stances can be non-continuous and may take place along several dimensions of behavior. If these dimensions of behavior can be aggregated to one perceived measure of norm deviation our model is applicable.

but never self-reported non-homophobic heterosexuals. At the same time, there were subjects stating to be non-homophobic heterosexuals, and they were all physically measured to be heterosexual. Another supportive evidence for this pattern is given by Weinstein et al. (2012), who show that disproportionately many of those claiming to be homophobic have homosexual tendencies. Additional casual observations, of religious leaders or members of conservative parties first supporting an anti-gay agenda and then caught conducting homosexual activities, are readily available to anyone surfing the Internet. In these cases it is clear that those homosexuals who publicly declare homophobic views, put pressure on others, and thereby take part in upholding a norm that seems disadvantageous to themselves. In our model this pattern may occur if individuals perceive a sufficiently concave disutility from misrepresentation. We do not know whether this assumption is valid when it comes sexual orientation. Our model simply predicts that if this would be the case, then inversion could arise as was observed. More precisely, it says this should happen in societies and groups upholding a conservative norm towards homosexuals but not in societies where homophobia is not normative (in which case homosexuals may instead sometimes state to be non-homophobic heterosexuals).[38,39]

A domain where it may be reasonable to expect people to have very concave preferences, is that of religious beliefs. That is, it seems plausible that if a person cannot follow her privately preferred religion, then she will be more or less indifferent about which other religion to profess. This means that if a certain religious group is being persecuted, members of that group will be likely to convert to the dominant religion, even in the presence of a third religion that is closer in religious terms and that is less persecuted. This will constitute an inversion of preferences if, in addition, the third religion is closer (in religious terms) to the dominant religion yet its members do not convert. This description seems to be illustrative of the mass conversion of Jews to Christianity under the Spanish inquisition (1391-1492

[38] A related observation, which will not be discussed here, relates to women's stated sexual preferences vis-à-vis their actual sexual arousal (Morokoff, 1985).

[39] In the psychological literature, the term *reaction formation*, one of Freud's classical defense mechanisms, has some resemblance to our term of inversion of preferences. There are, however, differences along a few dimensions. Firstly, reaction formation has the focus on an individual dealing with her own tastes vis-à-vis her thoughts of what is right. Secondly, reaction formation seldom compares individuals with different tastes, while such comparison is at the heart of our concept of inversion. Finally, as a consequence of the focus on the individual's interactions with herself, reaction formation theory is silent on what norms and pressure are bound to arise in a society. For a description and empirical observations of reaction formation see Baumeister et al (1998).

AD), with the "third" religion being Islam. During this period, the Christian rulers persecuted Jews fiercely while the Muslim population was persecuted considerably less (Baer, 1965, p.286;[40] Ruiz, 2008, p.160). While religious preference is probably a multidimensional object, the consensus among historians (see Ben-Shalom 2001, p.252; Grossman 1998, pp.30-34; and Ben-Sasson 1990, p.20) is that for a Jew, in that time and place, converting to Islam would have been a substantially smaller step than converting to Christianity.[41] So Jews could have lowered the pressure by converting to Islam, but instead converted to Christianity or stayed with their original faith, while the conversion of Jews to Islam is essentially unheard of (Baer, 1965). At the same time, Muslims (unlike Jews) accepted central aspects of Christianity such as the divinity of Christ and his second coming, the means of salvation and God's role in it, and views on afterlife. Yet Muslims did not convert to Christianity during this historical period (Ruiz, 2008, p. 160). The total pattern looks like an inversion of preferences, where those with religious views close to the Christian norm (Muslims) kept their true faith in public, while those with more distant religious views (Jews) converted in masses to Christianity. Their conversion most likely made it harder for the remaining Jews to be openly Jewish. So the converted Jews passively strengthened a norm that was disadvantageous to themselves. Some previously Jewish individuals were even actively persecuting non-converted Jews. For instance, Paul de Burgos and Geronimo de Santa Fe, both Jews who had converted to Christianity, took an active role both in tightening the anti-Jewish laws and in making Jews convert. This suggests the mechanism of our model was present during the Spanish inquisition.[42]

[40] The page number is given for the original Hebrew version of the book.

[41] For instance, in 15th century Spain, Jews and Muslims practiced polygamy and had similar religious bans (like the taboo on eating pork). Furthermore, among Jews, Christianity was often considered to be almost polytheistic. Many of these similarities and differences have persisted until today.

[42] We cannot rule out other mechanisms being present as well. It may, for instance, be that partial conversion would not have been acceptable. That is, it is possible Muslims of Jewish descent would not have been treated as Muslims but as unconverted Jews. This alternative description may seem reasonable only if the persecutors were able to track individuals and their religion over time. Likewise, we cannot rule out that Jews thought the Muslims would be next in line for persecution, thus considered it useless to convert to Islam. Had they correct expectations about this, they need not have worried personally, as persecution of Muslims affected only later generations.

6 Revolutions and mass protests

Let us now suggest a slightly different interpretation of the model by letting it represent political opinions and pressures. Then $\bar{s}$, being a cluster of individuals declaring the same political view, can be interpreted as a political regime. In a way, our model provides a generalization of previous, binary, models of regime support and collapse (e.g., Granovetter, 1978; Kuran, 1989a) which typically analyze social protests, riots and revolutions. In the binary models, individuals have the choice between supporting the regime or joining a protest movement. Each individual has a different propensity for each one of these alternatives and the propensity for each alternative increases the more other people choose that alternative. In the binary model, those who like the regime the most follow it while those who dislike it the most may dissent. The extension that our model brings in is by letting individuals choose *the extent* of support for the regime. They can fully support it, by stating $\bar{s}$, or choose any level of critique $s \neq \bar{s}$. This seems to be a plausible assumption. Likewise, it seems plausible that not only the number of regime supporters should determine its strength (like in the binary models), but also how critical those who publicly disagree with the regime are. Our static analysis then describes which individuals (i.e., having which private preferences) will uphold a regime, who the non-supporters will be and what they will do. In particular, it shows the conditions for the very existence of a political regime or cluster in equilibrium. Our dynamic analysis can be used to discuss what may spark the undermining of political regimes and how the ensuing process will look like. It answers questions (to the best of our knowledge previously unaddressed by formal models) such as: which private opinions those first out to protest will have and what they will publicly declare; who the next ones out will be; at what point opposite factions to the initial protesters will stop supporting the regime; and what the new equilibrium reached after the collapse will be.

6.1 Static analysis of political regimes

Starting with the static analysis, our model shows that privately disliked (i.e., biased) political regimes may indeed exist and display a large but fake public support. This holds even without the existence of an elite with coherent interests who supports the regime.[43] We will now discuss the insights for each type of society in this political

[43] Naturally, if a strong group with coherent interests exists, this can lead to additional clustering of supporters beyond that group.

setting.

Proposition 1 covers the case where individual political pressure is convex. Here, public stances will not at all cluster. One interpretation of this result is that no publicly supported political regime can exist in this setting. Alternatively it can be interpreted as inexistence of political cohesion. Of course, political pressure will still exist, arising from the individual statements in equilibrium. There may also exist some mainstream political opinion that minimizes political pressures. However, this mainstream opinion will not actually be stated by many in equilibrium, so we do not consider it to be a political regime.[44]

For a publicly supported regime to exist, political pressure has to be concave. For instance, in the alienating society a regime can be sustained by the population (see Proposition 2). Here those who like the regime the most follow it, while those who dislike it dissent.[45] We further show that in this equilibrium the dissent is not at all homogenous. Individuals will not form clusters of critique as dissenters will speak their (different) minds. Our analysis of the alienating society further shows that biased regimes may exist, but are weaker and essentially require a very high degree of initial conformity.

When society is of the inverting type, we get very different patterns than those of the alienating society. Here the political regime is upheld by those who dislike it the most, and those who do speak out against the regime do it in a very mild manner and are privately rather content with it (see Proposition 4). Furthermore, the critique essentially always comes from both sides of the regime. In this case, privately despised political regimes not only can exist, but are also more sustainable than more central regimes. This may explain why some political regimes seem to advocate policies so far from what one would think is in the people's interest, yet meet no extreme public objections.

6.2 Dynamic analysis of revolutions

We turn now to applying the dynamic analysis to the political setting. As a first intuitive result, it is clear that if the sanctioning power falls or people start valuing

[44] In Appendix A we further show that if all individuals speak their minds, then a mainstream opinion that minimizes political pressure exists and equals the average private opinion in society.

[45] This equilibrium is, in this regard, similar to the one described by Granovetter (1978) and Kuran (1989a) but is more elaborate. Firstly, this case is only one out of several possible cases. Secondly, our model yields a richer description of the dynamic change of stances and describes which regimes can be upheld in equilibrium.

their own individual tastes more (i.e., K falls), the political regime may collapse in both the alienating and inverting societies. But the more interesting insights come from considering shifts of opinions.

We begin our analysis by focusing on alienating societies (see Proposition 3). Here, gradual collapse will be preceded by a shift of private opinions away from the regime. This is depicted in Figure 5. There we start (top schedule) with a regime at $\bar{s} = 0$ and a type distribution between -1 and 1. Suppose that the type distribution moves gradually to the right. That is, private sentiments become less in line with the current regime. If K is sufficiently large, this gradual shift may be invisible on the surface, as all or most still conform to the political regime. This way *public* sentiments may not be affected and the political regime may stay at $\bar{s} = 0$ even though it no longer represents the average private opinion. This is depicted in the second schedule from the top. But after the type distribution shifts beyond a certain point (third schedule), opposition may arise and voice its discontent with the regime. This happens if K is not sufficiently large to uphold the old regime ($\bar{s} = 0$), which is now very biased compared to the new average private opinion. This opposition will be fierce in the sense that it is made of those who dislike the regime the most. Furthermore, these individuals will display no compromise when voicing their opinions and will not cluster. This may be a point of no return for the regime. Even if opinions cease to shift further, the opposition is bound to grow, because in the next period less extreme types will also be inclined to voice their private opinions, thus raising more (though milder) opposition (fourth schedule). This way the political regime will be undermined, until all speak their minds (bottom schedule). It will then become observable that the initial regime was actually no longer representative of the private opinions.[46] More generally, in this case the revolution will start only on one side of the political spectrum, say on the right, but at an intermediate stage also those on the far left of the regime will stop supporting it and instead speak their minds. The last ones to leave the regime will not be the most leftist, but those on both sides who closely agree with the regime. Our model also provides a clear prediction about the post collapse state – society will converge to pluralism. Moreover, this state will be absorbing, as further changes to the

[46] To some degree the description of the alienating society, up to here, is similar to the one described by Granovetter (1978) and Kuran (1989a). However, this description holds only when initially there is full cohesion in society. Another scenario, which is not depicted in the figure, is one where there is opposition already in the initial steady state. Then, gradual shifts of private preferences will induce also a gradual shift to less cohesion.

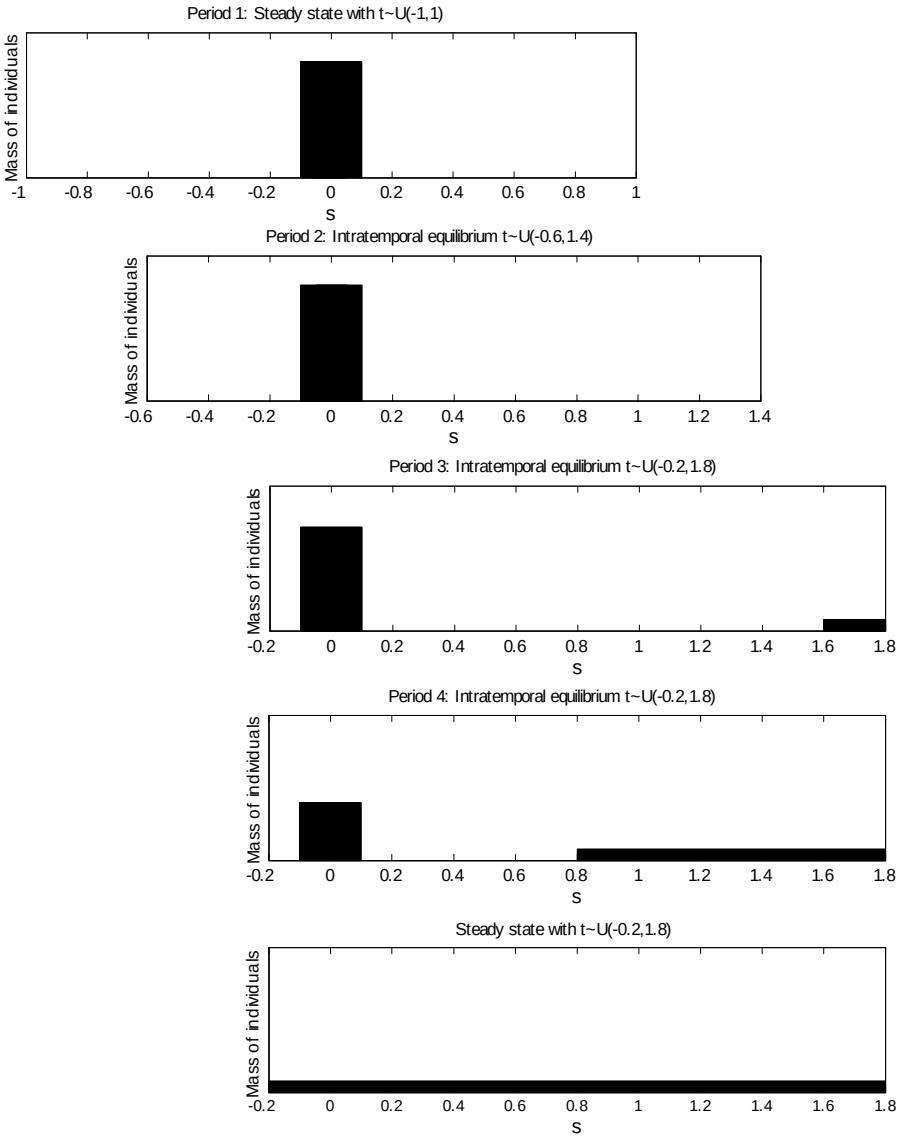

Figure 5: The distribution of stances over time when private sentiments change. $K = 1.3$, $\beta = 0.01$ and $\alpha = 0.5$. The width and placement of each horizontal axis represents the distribution of types in that period. Note that the bar representing the norm at $\bar{s} = 0$ should, strictly speaking, be infinitely narrow. But for clarity of exposition we depict it as wide. In the first schedule there is a single norm equilibrium where all types, $t \sim U(-1, 1)$, follow a political regime at $\bar{s} = 0$. In the second schedule the distribution types has changed to $t \sim U(-0.6, 1.4)$ yet all follow the regime. In the third schedule the type distribution has shifted further to $t \sim U(-0.2, 1.8)$ and some opposition arises. In the fourth schedule the type distribution has not changed further, but there is more opposition to the regime. In the fifth schedule the type distribution is unchanged but the political regime has collapsed and a new steady state has been reached, where are all types speak their minds.

type distribution will only lead to individuals declaring a new set of private opinions that reflect this new distribution. However, as often happens in real life, if there exists a (possibly strong) group with coherent private opinions, then a new regime may be established.[47] The overall pattern of revolutions in alienating societies, as described above, seems to provide a reasonable description of the Iranian revolution in 1978-79. There the regime became less aligned with the religious sentiments in society over time. The revolution was then initiated by the hardest opponents of the old regime but then gained mass support by recruiting more moderate individuals (Razi, 1987).

Focusing now on the collapse of regimes in inverting societies, we start once again by considering what would happen to the regime if private opinions in society changed to be less in line with it. This is depicted in the left part of Figure 6. Suppose we start in a stable equilibrium with a biased regime at $\bar{s} = -0.8$, while the distribution of types is between -1 and 1 (upper left part). If private opinions drift away from the regime to be between, say, -0.9 and 1.1 (lower left part), then there will be fewer types on the left of the political regime, implying that the political regime will only become stronger.[48] This is through the direct effect of there existing less people on the left and the indirect effect whereby it induces more conformity on the right as well. It can be seen by the right tail being shorter after the shift. Hence, unlike the previous society, here regimes do not break following private sentiments shifting away from it. This suggests that a regime or a religion that have been determined in some far away history may appear very strong today even if the private sentiments have shifted away.

However, focusing on the right part of the figure, now the regime may paradoxically break following a shift whereby private sentiments become *more in line* with the regime (alternatively, we can think of the regime's policies shifting to be more representative of the population). For the purpose of this discussion, suppose that K is just large enough to ensure that the regime at $\bar{s} = -0.8$ is stable when the private opinions are between -1 and 1 (top right schedule). Suppose now that private opinions gradually shift leftward. Then we will first see an increase of mild critique from

[47] The existence of such a group with coherent interests is assumed exogenously by Granovetter (1978) and Kuran (1989a). This assumption together with the constraint that individuals can only support the regime or this opposing group implies that in their model a succesful revolution will always lead to the establishing of a new regime rather than to a state of pluralism.

[48] Note that the dynamic analysis in Section 5 ensures that if the change is gradual enough and we start from a stable steady state, there *will* be convergence to a new equilibrium around the same norm, with a greater share of norm followers.

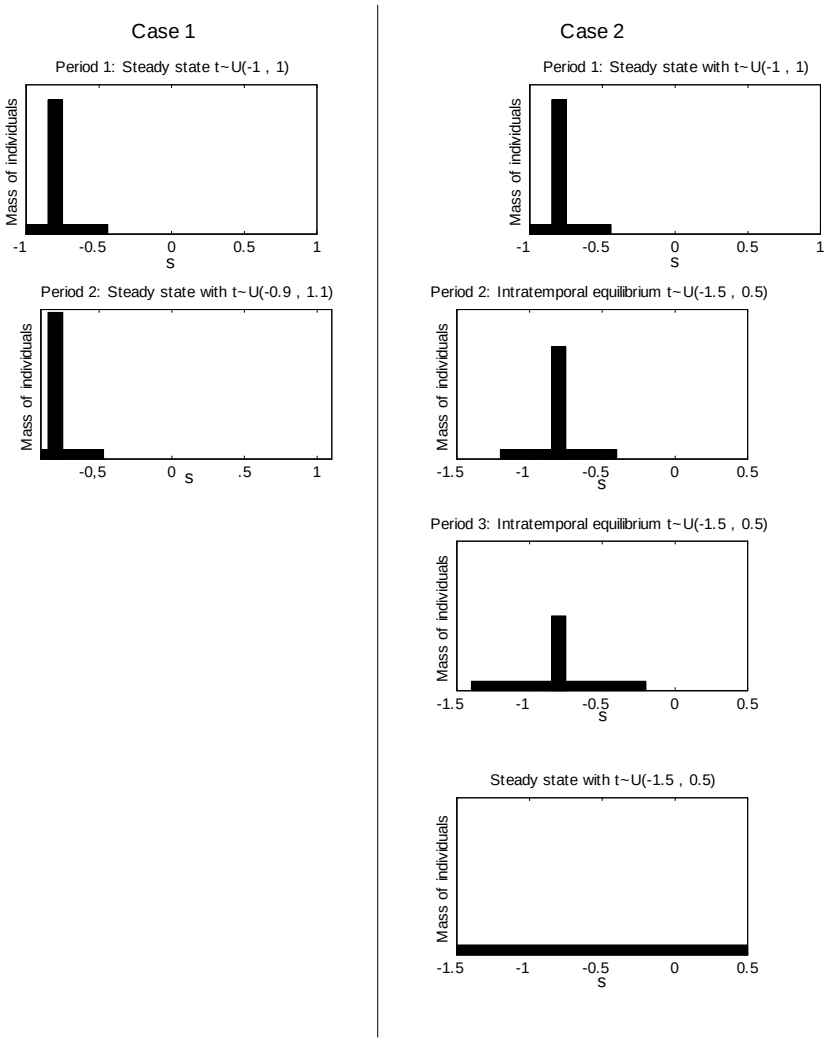

Figure 6: The distribution of stances over time when private sentiments change. $K = 2.1$, $\beta = .5$ and $\alpha = 0.01$. The width and placement of each horizontal axis represents the distribution of types in that period. Note that the bar representing the norm at $\bar{s}$ should, strictly speaking, be infinitely narrow. But for clarity of exposition we depict it as wide. The top schedule (both left and right) depicts a single norm equilibrium under a uniform distribution of types in $[-1, 1]$, where some follow a political regime at $\bar{s} = -0.8$. Case 1 (on the left): in the second schedule the distribution of types has changed to $t \sim U(-0.9, 1.1)$, and as a result more individuals follow the political regime. Case 2 (on the right): In the second schedule the type distribution has shifted to $t \sim U(-1.5, 0.5)$, and more individuals speak their minds. In the third schedule the type distribution has not changed further, but there is more opposition to the regime. In the fourth schedule the type distribution is unchanged, but the political regime has collapsed and a new stable steady state has been reached, where are all types speak their minds.

the left. This will spur more mild critique also from the right. This is since speaking one's mind becomes more appealing to people on the right when the dissidents on the left reduce their support of the regime (second schedule from top). These two effects will work to enhance each other, gradually stretching the borders of free expression in public, and the regime will be followed by fewer and fewer (third schedule from top). Since K is by assumption too small to keep upholding a regime at $\bar{s} = -0.8$, this will lead to the regime gradually collapsing (bottom schedule). Society will now settle on a new absorbing state of pluralism, where there exists no regime or cluster of public opinions. If the distribution of types would shift further from this point there would be an equivalent shift in stated opinions.

The fundamental difference from the dynamics of the alienating society (and binary models of revolutions) is that now regimes will break, not by fierce opposition, but rather following a process where initialy the criticque is mild but then gradually intensifies on both sides of the political spectrum. An interpretation of this is that now a political regime is undermined from the inside out, by internal opposition legitimizing more and more extreme views, rather than by external force that increasingly gains the popularity of those with less extreme views.

The revolutionary pattern of the inverting society is particularly appealing as it provides a theoretically consistent explanation for an important class of mass revolutions which were previously unexplained by formal theory. For example, the inside-out pattern seems to be a reasonable description of the protest movements that led to the collapse of some of the communist regimes in Eastern Europe in 1989-90 and to the recent Arab Spring protests in Egypt. In Eastern Europe, the initial protests were not very extreme. For instance, Hungarian communist party leader Karoly Grosz expressed that "the party was shattered not by its opponent but – paradoxically – from within" (Przeworski 1991:56). Furthermore, in Poland and Hungary, moderate dissidents instigated liberal reforms and made demands for free elections (Pfaff, 2006). Similarly, as was reported about Egypt, the most extreme factions (i.e., the Muslim Brotherhood and the Salafi movement), were hardly present in the protests initially.[49]

[49]For instance, a BBC news profile on the Muslim Brotherhood reports that initially "(t)he group's traditional slogans were not seen in Cairo's Tahrir Square. But as the protests grew and the government began to offer concessions, including a promise by Mr Mubarak not to seek re-election in September 2011, Egypt's largest opposition force took a more assertive role". See http://www.bbc.com/news/world-middle-east-12313405.

Another important feature of the inverting society is that the undermining of the regime is initiated by individuals with opinions on *both* sides of the political spectrum. When it is undermined by a revolution "from the outside in", as in the alienating society, the revolution will always start at one end of the spectrum and eventually spill over to the other end. However here, when the regime is undermined by mild critique, the whole process starts with voiced critique on both sides of the regime (unless the regime is so biased that there are opinions on only one of the sides). This way we get that regimes may be undermined by truly "strange bedfellows", in the sense that they are pulling the public opinion space in two different directions. This was a clear pattern in the Arab spring revolution in Egypt. The protesters on the Tahrir square consisted of some who suggested that Mubarak was not sufficiently liberal and of others who expressed that he was not conservative enough. While the spark may have been a shift in private opinions towards more liberalism (a leftward movement of the opinion axis when moving from right hand schedule 1 to right hand schedule 2 in Figure 6), the later elections showed that in fact Egyptian society was more conservative than Mubarak's regime (in line with the description Figure 6, where the average opinion after the shift is to the right of $\bar{s} = -0.8$, which represents Mubarak's regime in that figure). This way, our analysis predicts that in revolutions that go from the inside out, what may initially seem to be a leftist revolution ends up being a rightist revolution instead.

The model also has something to say about failed revolutions. Sometimes a popular protest seems to gain initial support that stops increasing at some point, eventually failing to topple the regime. In the model this can happen under two main scenarios. The first scenario occurs when a shift in private sentiments reduces support for the regime but K is sufficiently large to uphold a regime also after some supporters have left it. For instance, on the right hand side of Figure 6, a change in private sentiments implies going from schedule 1 to 2 and further on to 3. An increase in public critique is then observed. This could have been a sign of imminent regime collapse. But if K is large enough, this process need not actually lead to collapse. Rather, the process could halt at schedule 3 where the regime is still in power though with less public support.

The other scenario is when initially there exist two equilibria of regime support (i.e., the same $\bar{s}$ may support two steady states with different degrees of conformity), but after a shift of private opinions or a weakening in regime strength (a lowering

of K) only one of the equilibria remains. Then, if society started in the equilibrium with the higher degree of conformity and ended up in the one with lower degree of conformity (because the former stopped being an equilibrium), a failed revolution will be observed. This scenario can happen when the political regime is biased in a way that enables the revolution to gain support only from one side of the political spectrum. This constraint on the gained support has different implications on regime fragility in each type of society. In an alienating society with a biased political regime, the initial protesters are always on one side only. In a failed revolution this one sided protesting will not gain the sufficient momentum needed to eventually induce protests on the other side too, as is required for a complete collapse.[50] This means that in an alienating society a revolution will be especially fragile in the beginning, when it only adds dissidents on the same side. But once individuals on the other political extreme start leaving the regime too, the revolution is bound to succeed. In an inverting society we get the opposite result with respect to revolutionary fragility. In the beginning of a revolution the critique intensifies on both sides of the regime. But it is when one side has been exhausted (so that there are no more extreme individuals on that side) that the revolutionary momentum falls, as it now has to rely on additions from one side only. At this stage the revolutionary process may halt, resulting in a regime that has been weakened yet not collapsed.[51]

7 Conclusions

This paper studies the sustainability and dynamic stability of endogenous social norms and political regimes under peer pressure. In many situations characterized by peer pressure, individuals may truly disagree (on a private level) about the right ideology

[50] This can be seen in equation (12), which is less steep when y is large. Graphically it can be seen by noting that in Figure 2 there is a kink of the dynamic function (this kink exists for all values of $\bar{s}$ except for zero and one). A shift of private sentiments or a lowering of K implies tilting the curve downwards, which with the kink can imply going from a full regime support steady state to one with mixed support.

[51] Moreover, if we relax the assumption of a uniform distribution of types, failed revolutions can occur also due to the form of the type distribution. Basically, the mechanism would resemble that of a failed revolution in a binary model. In a binary model, the revolution fails if, after gaining a certain amount of support, the next ones to join it are reluctant to do so because the existing revolutionary support is not enough to trigger them. Similarly, if in our model the distribution is not uniform and the mass of individuals is too low at some range, then once the next ones to join the revolution are supposed to be recruited from this exact range, the acceleration will slow down and the revolution will fail.

or best conduct. Hence, there will not exist a consensus opinion that can make for an exogenous norm. Nevertheless, we show that in these situations a clear norm may be endogenously sustained and will also be dynamically stable. That is, there may seem to exist a consensus opinion that many adhere to, when in fact individual preferences are completely heterogeneous. Moreover, it will often be the case that a norm which is biased with respect to private preferences will be more sustainable than a representative norm. This can shed light on the sustainability of biased norms, as observed in religious communities, racial attitudes, honor cultures and various autocratic societies.

The paper maps societies into a class that cannot maintain an endogenous norm and a class that can. Within the class that *can*, the paper highlights a fundamental difference between two main subclasses of societies. Firstly, in societies where pressure is sufficiently concave, individuals with opinions that are very different from the norm will be *alienated* and state those private opinions in public. For a norm to survive in this type of society, it has to be sufficiently centrally located and to closely represent the opinion of most individuals in society. If society is either very heterogenous, or the norm is biased, a norm can be sustained only under strong pressure. In the other subclass of societies, where pressure is not sufficiently concave, preferences will be *inverted* – the ones speaking their minds openly will be those with private opinions that are only slightly different from the norm, while those who privately dislike the norm the most will fully conform. This means that in this type of society we should observe mild critique of the norm (which can be interpreted as an internal opposition). Here biased norms are more sustainable and more magnetic than central norms.

The model can also be interpreted as being about the formation and collapse of political regimes. Naturally, if we would assume the existence of a group with aligned interests, political pressure could lead to additional clustering of opinions by people beyond that group. But what we show is that a regime can be sustained even in the absence of such a group, i.e., even when private interests are fully heterogeneous. Under this interpretation, the model explains the existence of biased regimes that are publicly supported. The dynamic analysis also highlights what we should expect to be the spark leading to the undermining of a regime in each society. In an alienating society, if a regime is to be undermined, this should be expected to happen through a process of fierce opposition. This opposition will arise at one end of the spectrum, as a consequence of private opinions having moved away from the regime, possibly

without detection. In an inverting society on the other hand, a collapse of the regime will be initiated by private sentiments becoming more in line with the regime. This will increase the amount of public criticism, which will be initially mild and will come from both sides of the political spectrum. The further evolution will then be a gradual stretching of the freedom of speech to include more critical statements. Hence, in inverting societies revolutions will go from the inside towards the outside.

We believe that the model in this paper represents an essential element in human interaction. Namely, that peer pressure arises in between multiple individuals. While some norms may be institutionalized, there are many situations where a norm would not exist unless (sufficiently many) individuals actually followed it. But even in situations where a norm or a regime *is* institutionalized, it seems reasonable that the extent of conformity to it, and what non-conformers do, should take part in determining the strength of the norm or the regime. Analytically proving outcomes in this setting is not a trivial matter and we have not exhausted the possible equilibria that can arise. However, our results of the dynamic model strongly indicate that the single norm equilibrium, which has been the focus of this paper, is not just a technical possibility – outcomes will tend to gravitate towards this kind of equilibrium from a broad set of initial conditions.

References

[1] Acemoglu, D., & Jackson, M. O. (2011). "History, expectations, and leadership in the evolution of social norms" Working paper No. w17066, National Bureau of Economic Research.

[2] Acemoglu, D., & Jackson, M. O. (2014). "Social Norms and the Enforcement of Laws" Working paper No. w20369. National Bureau of Economic Research.

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

[4] Adams, H. E., Wright, L. W. Jr. and Lohr, B. A., (1996), "Is Homophobia Associated With Homosexual Arousal?," *J. of Abnormal Psychology*, Vol. 105, No. 3, pp. 440-445.

[5] Arendt, H. (1964). *Eichmann in Jerusalem*. New York: Penguin Books.

[6] Akerlof, G. A. (1997). "Social distance and social decisions". *Econometrica*, Vol. 65, No. 5, pp. 1005-1027.

[7] Almer, C., Laurent-Lucchetti, J., Oeschlin, M. (2013). "Income shocks and social unrest: theory and evidence". mimeo Tilburg University.

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

75(3), 711-756.

[9] Baer, Y., (1965), "A history of the Jews in Christian Spain"; translated from the Hebrew by Louis Schoffman. Philadelphia: Jewish Publication Society of America, third edition.

[10] Baumeister, R. F., Dale, K. and Sommer, K. L., (1998), "Freudian Defense Mechanisms and Empirical Findings in Modern Social Psychology: Reaction Formation, Projection, Displacement, Undoing, Isolation, Sublimation, and Denial," *J. of Personality*, Vol. 66, No. 6, pp. 1081-1124.

[11] Beck, C. J. (2009). *Ideological Roots of Waves of Revolution.* ProQuest.

[12] Bénabou, R. & Tirole, J., (2006), "Incentives and Prosocial Behavior", *American Economic Rev*, 96(5), 1652-1678

[13] Ben-Shalom, R., (2001), "Kiddush Hashem and Jewish Martyrology in Aragon and Castile in the Year 1391: Between Spain and Ashkenaz," *Tarbitz*, Vol. 70, No. 2, pp. 227-282. [in Hebrew]

[14] Ben-Sasson, M., (1990), "To the Jewish Identity of the Anusim: an Advisement in the Hishtamdut at the Period of the Almohad Caliphate," *Peamim*, Vol. 42, pp. 16-37. [in Hebrew]

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

[16] Bisin, A., & Verdier, T. (2001). "The economics of cultural transmission and the dynamics of preferences". *J. of Economic Theory*, 97(2), 298-319.

[17] M. Blumenthal, C. Christian, and J. Slemrod. (2001) "Do Normative Appeals affect Tax Compliance? Evidence from a Controlled Experiment in Minnesota". *National Tax J.*, 54(1):125-138.

[18] Borsari, B., & Carey, K. B. (2001). "Peer influences on college drinking: A review of the research". *J. of substance abuse*, 13(4), 391-424.

[19] Bowles, S. (1998). "Endogenous preferences: The cultural consequences of markets and other economic institutions". *J. of economic literature*, Vol. 36, No. 1, pp. 75-111.

[20] Brock, W.A., Durlauf, S.N., (2001), "Discrete Choice with Social Interactions", *Rev. of Economic Studies* Vol. 68, pp. 235–260.

[21] Centola, D., Willer, R., & Macy, M. (2005). "The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms". *American J. of Sociology*, 110(4), 1009-1040.

[22] Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior". *Adv. in experimental social psych*, 24(20), 1-243.

[23] Cialdini, R. B. (2003). "Crafting normative messages to protect the environment". *Current directions in psychological science*, 12(4), 105-109.

[24] Clark, A. E., & Oswald, A. J. (1998). "Comparison-concave utility and following behaviour in social and economic settings." *J. of Public Economics*, 70, 133-155.

[25] Cohen, D. (2001). "Cultural variation: considerations and implications". *Psy-*

chological bulletin, 127(4), 451.

[26] Colson, E. (1975) *Tradition and contract: The problem of order*. Chicago: Aldine.

[27] D'Augelli, A. R., (2006), Developmental and contextual factors and mental health among lesbian, gay, and bisexual youths. In A. E. Omoto & H. M. Kurtzman (Eds.), Sexual orientation and mental health: Examining identity and development in lesbian, gay, and bisexual people (pp. 37–53). Washington, DC: APA Books. doi:10.1037/11261-002.

[28] Davis, J. A. (1959). "A formal interpretation of the theory of relative deprivation". *Sociometry*, Vol. 22, No. 4, pp. 280-296.

[29] Davies, J. C. (1962). "Toward a theory of revolution". *American sociological review*, Vol. 27, No. 1, pp. 5-19.

[30] Dufwenberg, M., & Lundholm, M. (2001). "Social norms and moral hazard". *The Economic J.*, 111(473), 506-525.

[31] Esteban, J. M., & Ray, D. (1994). "On the measurement of polarization". *Econometrica*, Vol. 62, No. 4, pp. 819-851.

[32] Esteban, J., & Ray, D. (2011). "A model of ethnic conflict". *J. of the European Economic Association*, 9(3), 496-521.

[33] Festinger, L. (1954). "A theory of social comparison processes". *Human relations*, 7(2), 117-140.

[34] Fields, J. M., & Schuman, H. (1976). "Public beliefs about the beliefs of the public". *Public Opinion Quarterly*, 40(4), 427-448.

[35] Gino, F., Norton, M. I., & Ariely, D. (2010). "The Counterfeit Self The Deceptive Costs of Faking It." *Psychological Science*, 21(5), 712-720.

[36] Gladwell, M. (2000). *The tipping point*. Boston: Little, Brown.

[37] Gneezy, U., Rockenbach, R., and Serra-Garcia, M. (2013), "Measuring lying aversion", *J. of Economic Behavior & Organization*, Vol 93, pp. 293–300.

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

[39] Grossman, A., (1998), "Kiddush Hashem in the 11th and 12th Centuries: Between Ashkenaz and the Muslim Countries," *Peamim*, Vol. 75, pp. 30-34.

[40] Herrmann, B., Thöni, C. and Gächter, S. (2008), "Antisocial Punishment Across Societies," *Science*, Vol. 319, pp. 1362–1367.

[41] Kandel E., Lazear, E. P., (1992), "Peer Pressure and Partnerships," *The J. of Political Economy*, Vol. 100, No. 4, pp. 801-817.

[42] Kendall, C., Nannicini, T., & Trebbi, F. (2013). "How do voters respond to information? Evidence from a randomized campaign" NBER WP 18986.

[43] Kitts, J. A. (2003). "Egocentric bias or information management? Selective disclosure and the social roots of norm misperception". *Social Psychology Quarterly*, 222-237.

[44] Kriesi, H., Koopmans, R., Duyvendak, J. W., & Giugni, M. G. (1992)."New social movements and political opportunities in Western Europe". *European J. of political research*, 22(2), 219-244.

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

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

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

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

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

[50] Lindbeck, A., Nyberg, S. and Weibull, J. W. (2003), "Social norms and Welfare State Dynamics", *J. of the European Economic Association*, Vol 1, Iss 2-3, pp. 533–542.

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

[52] Lopez-Pintado, D., & Watts, D. J. (2008). "Social influence, binary decisions and collective dynamics". *Rationality and Society*, 20(4), 399-443.

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

[54] McAdam, D., Tarrow, S., & Tilly, C. (2003). "Dynamics of contention", *Social Movement Studies*, 2(1), 99-102.

[55] McAdams, R. (1997). "The origin, development, and regulation of norms". *Michigan Law Rev.*, 96, 338-433.

[56] Merton, R. K., and A. S. Kitt, (1950), "Contributions to the Theory of Reference Group Behavior" in R. K. Merton and P. F. Lazarsfeld, Continuities in Social Research, Studies in the Scope and Method of "The American Soldier," Glencoe, Ill.: The Free Press, pp. 40-105.

[57] Michaeli, M. & Spiro, D., (2014),"The Distribution of Individual Conformity under Social Pressure across Societies". University of Oslo Dept of Economics working paper series Memo 12/2014

[58] Milgram, S. (1992). "The experience of living in cities". In S. Milgram (Ed.), The individual in a social world: Essays and experiments (pp. 10-30). New York: McGraw-Hill.

[59] Miller, D., & Prentice, D. (1994). "Collective errors and errors about the collective". *Personality and Social Psychology Bulletin*, 20, 541-550.

[60] Morokoff, P. J., (1985), "Effects of Sex Guilt, Repression, Sexual "Arousability," and Sexual Experience on Female Sexual Arousal During Erotica and Fantasy," *J. of Personality and Social Psychology*, Vol. 49, No. 1, pp. 177-187.

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

[62] O'gorman, H. J. (1975). "Pluralistic ignorance and white estimates of white support for racial segregation". *Public Opinion Quarterly*, 39(3), 313-330.
[63] Osborne, M. J. (1995). "Spatial models of political competition under plurality rule: A survey of some explanations of the number of candidates and the positions they take". *Canadian J. of Economics*, Vol. 28, No. 2, pp. 261-301.
[64] The Pew Research Center. (2007). World publics welcome global trade—but not immigration. Washington, DC.
[65] Pfaff, S. (2006). *Exit-voice Dynamics and the Collapse of East Germany: the Crisis of Leninism and the Revolution of 1989*. Duke university Press.
[66] Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America*. Cambridge University Press.
[67] Razi, G. H. (1987). "The Nexus of Legitimacy and Performance: The Lessons of the Iranian Revolution". *Comparative Politics*, 453-469.
[68] Robinson, C. E. (1932). *Straw votes*. New York: Columbia University Press.
[69] Roland, G. (2004). "Understanding institutional change: fast-moving and slow-moving institutions". *Studies in Comparative International Development*, 38(4), 109-131.
[70] Rubin, J. (2014). "Centralized institutions and cascades". *J. of Comparative Economics*.Vol 42, Iss 2, pp. 340–357
[71] Ruiz, T. F., (2008), *Spain's Centuries of Crisis: 1300-1474*. John Wiley & Sons.
[72] Savin-Williams, R. C., Ream, G. L., (2003), "Sex Variations in the Disclosure to Parents of Same-Sex Attractions," *J. of Family Psychology*, Vol. 17, No. 3, pp. 429–438.
[73] Schanck, R. L. (1932). "A study of a community and its groups and institutions conceived of as behaviors of individuals". *Psychological Monographs*, 43(2), i.
[74] Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams, Jr., (1949) *The American Soldier: Adjustment during Army Life*, Princeton, N. J.: Princeton University Press.
[75] Tanter, R., & Midlarsky, M. (1967). A theory of revolution. Journal of Conflict Resolution, 11(3), 264-280.
[76] Tarrow S. (1998), *Power in Movement*. New York: Cambridge U. Press. 2nd ed.
[77] Tullock, G. (1971). "The paradox of revolution". *Public Choice*, 11(1), 89-99.
[78] Vandello, J., & Cohen, D. (2000). "Endorsing, enforcing, or distorting? How southern norms about violence are perpetuated". Unpublished manuscript, Princeton University, Princeton, NJ.
[79] Weinstein, N., Ryan, W. S., DeHaan, C. R., Przybylski, A. K. , Legate, N. and Ryan,R. M., (2012), "Parental Autonomy Support and Discrepancies Between Implicit and Explicit Sexual Identities: Dynamics of Self-Acceptance and Defense," *J. of Personality and Social Psychology*, Vol. 102, No. 4, pp. 815–832.
[80] Wilson, J. Q., & Kelling, G. (1982). "Broken windows". *Atlantic*, 29-38.

A Appendix: Descriptive and prescriptive norms (for online publication)

The main focus thus far has been on norms that reflect statements that are actually made by a mass of individuals. In that sense we argued that the norms were descriptive. But there are other ways one could define a social norm. In particular, the main alternative to a descriptive norm (emphasizing what people actually do) is a prescriptive norm (emphasizing what people should do). While the former is rather straightforward, the latter is less so. In principle, prescriptive norms could fit situations where there is a consensus about what the right thing to do is, but where achieving this optimum is costly, as in models of status (e.g. Clark and Oswald, 1998) or of work effort (e.g. Kandel and Lazear, 1992). That would basically imply a social pressure that increases in the distance from an exogenously given norm. However, a broader attitude, which is pursued in this section, is to consider the stance that actually minimizes the aggregate social pressure P as reflecting what people approve and so find to be normative. This way, the definition of what one should do is neither arbitrary nor exogenous – it stems directly from the expectations of others, as reflected by the pressure imposed on different statements. The next definition then follows.

Definition 3 *A prescriptive norm is a statement $\tilde{s}$ that is a global minimum point of the social pressure P. Furthermore, if $\tilde{s} \neq \bar{s}$, or if $\nexists\, \bar{s}$, then $\tilde{s}$ is additionally called a virtual norm.*

This definition connects the prescriptive norm $\tilde{s}$ to the definition of a descriptive norm $\bar{s}$, while relating to the following typology. In a society there may exist a prescriptive norm that is descriptive too ($\tilde{s} = \bar{s}$). Alternatively, there may exist a prescriptive norm that is not descriptive, either because bunching happens elsewhere ($\tilde{s} \neq \bar{s}$) or because there is no descriptive norm at all ($\nexists\, \bar{s}$). It is in this last case that we call $\tilde{s}$ virtual, as it exists without anyone stating it. That is, there is no requirement that anyway would actually follow a prescriptive norm. This is indeed often the case when considering norms that reflect an ideal behavior, which no one can actually practice in reality. However, as we show in the following proposition, a prescriptive norm may be virtual even when it can be easily followed.

Proposition 6

1. *In the single norm equilibria described in Propositions 2 and 4, the descriptive social norm is also prescriptive ($\bar{s} = \tilde{s}$).*

2. *Let D be given by (5) and let p be given by (6) with $\beta > 0$, and consider an equilibrium in which all individuals speak their minds. Then there exists a prescriptive norm $\tilde{s}$ that is virtual and equals the average type.*

The proposition explores the existence of descriptive and prescriptive norms in all the equilibria discussed in the paper. The proof of the proposition is in the formal appendix, but the intuition is rather straightforward. The first statement follows directly from the fact that when p is a step function (Proposition 2), the pressure is reduced only where there is bunching, i.e., at the descriptive norm; and when D is a step function (Proposition 4), the pressure monotonically increases in the distance from the descriptive norm. But the lesson is more general. If there is one statement $\bar{s}$ that is made by many in society, then this will induce P to be rather low at $\bar{s}$. Conversely, if $\bar{s}$ was not the minimum point of P, then there would be no point in stating it. This suggests that in order to uphold a descriptive social norm $\bar{s}$, it needs to minimize social pressure, as otherwise there will be no bunching there.[52]

The second statement of the proposition states that in pluralistic societies, in which all individuals speak their minds, there will be a prescriptive norm even in the absence of clustering. This means that people will still feel peer pressure and that there can be a perceived consensus opinion even though (almost) no one actually states it.[53] Here the prescriptive norm is virtual since one can reduce pressure substantially by choosing a compromise solution in between full conformity on the one hand, and one's bliss point on the other hand. This norm will never be biased – its location will always equal the average private opinion in society, and so it is bound to be representative of the private sentiments in a pluralistic society.

B Appendix: Relaxing some model assumptions (for online publication)

The logic described thus far has implications beyond the case of a uniform distribution of types. When $\beta < \alpha \leq 1$, the more general lesson is that a single norm will tend to be accompanied by alienation of those who privately dislike the norm. This also has implications for the location of the norm and for the level of cohesion in society. It implies that unless K is very big, the norm can be sustained only if it is located such that most in society fairly like it privately. This is since otherwise there would be a large portion of opposers to the norm, who, by opposing, would make conforming

[52] This can also be related to the dynamic version of the model. For there to be convergence to a single norm equilibrium at $\bar{s}$, at each stage of the dynamic there have to be sufficiently many who state it, thus making it a prescriptive norm at that stage. Otherwise the norm would loose its magnetic power. Hence, requiring that $\bar{s} = \tilde{s}$ in the dynamic equilibrium both intra- and intertemporally is a necessary condition for the convergence to and maintenance of that equilibrium. However, it is not sufficient, since $\bar{s}$ also has to be sufficiently good at lowering social pressure for it to become a focal point of attraction. This can be seen in the phase diagrams. In Figure 2, x_{conv} marks the minimum degree of initial conformity that induces dynamic convergence in the alienating society. This is despite the fact that $\bar{s}$ would equal $\tilde{s}$ even if $x_i < x_{conv}$. A similar description applies to the inverting society: in Figure 4, y_{uss} marks the border of convergence although there exist $y_i > y_{uss}$ whereby $\bar{s}$ is still a prescriptive norm (but is not strong enough to induce convergence).

[53] The only case where this is not true is when p is a step function, in which case P is constant and independent of s, implying also that there is no unique minimum point of social pressure.

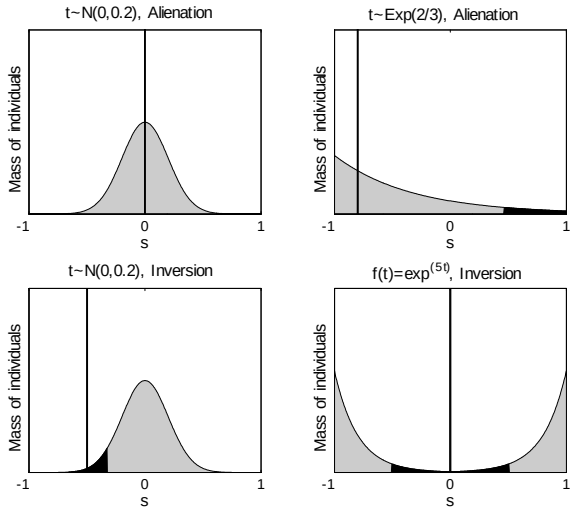

Figure 7: Histograms with single norm steady states in the dynamic model. The black surfaces represent the steady state distribution of stances while the grey surfaces represent the underlying distribution of types. Stance distribution in the zeroth generation is such that all state the same opinion. Note that the y-axis has been truncated for visibility and that the distributions of types has (when applicable) been truncated to be between -1 and 1. Upper left: $\alpha = 0.5$, $\beta = 0.01$, $K = 1.2$, $\bar{s} = 0$. Upper right: $\alpha = 0.5$, $\beta = 0.01$, $K = 1.2$, $\bar{s} = -0.8$, the distribution is exponential with mode at -1. Lower left: $\alpha = 0.01$, $\beta = 0.5$, $K = 2.5$, $\bar{s} = -0.5$. Lower right: $\alpha = 0.01$, $\beta = 0.5$, $K = 1.5$, $\bar{s} = 0$.

unattractive even for those who object the norm less. Figure 7 shows steady states under some other distributions of types. Under a normal distribution, the norm has to be located within the bell of the normal distribution (upper left in Figure 7). Or, if the whole distribution is skewed, the norm needs to be located at the same side as the mass of types (upper right).

When $\alpha < \beta \leq 1$, the more general lesson is that single norms will be accompanied by inversion of preferences virtually regardless of the distribution of types. Furthermore, for a norm to be sustainable and have a high degree of cohesion it has to be located *away* from any mass of private opinions. Otherwise, if there is a mass of people with opinions close to the norm, these people will choose to speak their minds, and by doing so will make the norm less attractive even for those whose opinions are further away (and are therefore subject to more pressure when speaking their minds). This can be seen in Figure 7 (bottom left), where we illustrate a case with a normal distribution of types. The norm cannot be sustained within the bell-shape but only at the tails. On the bottom right of Figure 7 we see that if the type distribution is

bimodal, a norm can be sustained virtually anywhere except close to the peaks.[54]

Beyond the assumption of a uniform distribution of types, which is not crucial as just explained, we have also implicitly assumed stationarity in our dynamic analyses (with the exception of the gradual shift in the type distribution discussed in the previous section). When interpreting our basic dynamic model as representing the way people update their stances given new information about others' stances, this assumption is innocuous. However, if one wishes to interpret the model as an overlapping generations model, stationarity may seem a strong assumption, and some other alternative ways to model come to mind. In particular, one may expect the type distribution to change following the statements made in previous generations. This change can be determined either by an exogenous rule (as in Kuran and Sandholm, 2008) or by an endogenous decision made by the individuals in the previous generation (as, e.g., in Bisin and Verdier, 2001; for a broader discussion of endogenous preferences see Bowles, 1998, and for a discussion of slow and fast moving institutions see Roland, 2004).

Starting with exogenous rules, a first option would be to let the distribution of types in one generation equal the distribution of *stances* taken in the previous generation. This could represent the case where each child is born with private preferences equaling the stated preference of her parent. Alternatively, one could interpret this as the parent making a decision how to raise the child (i.e., the parent is choosing s), taking into account the parent's private opinion t and the stated opinions in society. At any rate, this way of modeling would imply identical convergence to a norm equilibrium (since each individual chooses between declaring her type and the norm, and the social pressure only depends on stated opinions), but it would make collapsing of norms harder to obtain, as private preferences would gradually become more in line with the norm. A second option is that of Kuran and Sandholm (2008), where the private preferences of the children equal the average of the parents' statements and types. Now, whether our results will be replicated under this alternative depends on the more detailed modeling of pressure. Kuran and Sandholm (2008) assume that each generation creates its own equilibrium, taking into consideration only the individuals of that generation. This is fine under their assumption of double quadratic functions as the static equilibrium is always unique. But in our setting, where the static equilibrium is not unique, this would mean the dynamic process may look in many irregular ways.[55]

[54] Under a skewed distribution (e.g. an exponential distribution) it may be possible to support a norm also close to the peak of private preferences. However, this norm will only be followed by very few types with private opinions in the tail, while the vast majority will speak their minds. Hence, the simulation and intuition based conjecture is that it will be hard to sustain a norm with *high degree of cohesion* if the norm is placed where many agree with it fairly much but not fully.

[55] Unless we let children feel pressure from the whole parent generation (apart from inheriting their average stated and private preferences). If statements made by the parents put pressure on all kids, some ground can be gained. In the case of an alienating society the dynamic process should not be affected (although, like before, the collapse of norms becomes less likely). This is since all

As for the dynamic endogenous preference structure of Bisin and Verdier (2001), it is harder to apply it to our setting. This is since their model contains only binary preferences. The parent of type A chooses how much effort to exert to make the child grow to be of type A too. But with some probability (which depends on the mass of each type in society), the child may become of type B. Now, to the best of our knowledge, no one has analyzed the possibility of the child becoming a type on a continuum between A and B. Probably such an addition would change or enrich the results (for the same reason that the binary model results of Kuran 1989a and Granovetter 1978 are enriched by letting stances be chosen from a continuum like we let them). But it is not obvious how to do that, as this would fundamentally change the decision problem of the parent.

C Appendix: Proofs (for online publication)

C.1 Initial results

Lemma 5 *Let there be a range of types that speak their minds. Then the aggregate pressure that results is strictly increasing in the distance from the middle of the range.*
Proof. *Let the distribution of stances be uniform at* $[a,b]$ *with* $a<b$. *Then*

$$P(s)=\frac{1}{2}K\int_{a}^{b}|s-\tau|^{\beta}\,d\tau$$
$$=\begin{cases}\frac{1}{2}K\frac{(a-s)^{\beta+1}-(b-s)^{\beta+1}}{\beta+1}<0 & \text{if } s<a\\ \frac{1}{2}K\frac{(s-a)^{\beta+1}-(b-s)^{\beta+1}}{\beta+1} & \text{if } a\leq s\leq b\\ \frac{1}{2}K\frac{(s-a)^{\beta+1}-(s-b)^{\beta+1}}{\beta+1}>0 & \text{if } s>b\end{cases}$$

It is easy to see that $P'(s)>0$ *if* $s>\frac{a+b}{2}$ *and* $P'(s)<0$ *if* $s<\frac{a+b}{2}$, *implying that* $P(s)$ *is strictly increasing in the distance from the middle of the range.* ■

C.1.1 Proof of Proposition 1

First we note that there cannot be a norm at one of the distribution edges, i.e., at $\bar{s}=-1$ or at $\bar{s}=1$. To see this, note that a norm at $\bar{s}=1$ for example implies that the slope at the norm is positive (because deviation to the left decreases the pressure from all stances besides $s=1$ while, when $\beta>1$, not affecting the pressure stemming from the mass of people at the norm), and so everyone would like to deviate to the left,

types within a certain range fully conform. So when a parent of type t conforms by stating $s=\bar{s}$, the child will become of type $(\bar{s}+t)/2$, which is closer to the norm than t is. This implies also that the child will choose to conform. In the case of an inverting society, the convergence may no longer hold. When all types close to the norm speak their minds, the child of a conforming extremist will typically be born as a moderate, who will then choose to speak her mind. This should lead to cohesion of private preferences but also to a gradual disappearance of the norm itself.

contradicting the existence of a norm there. Next we consider norms at the interior of $[-1,1]$. Here, note that when $\beta > 1$ then p and p' are continuous everywhere, which implies that $P = \int p$ and $P' = \int p'$ must be continuous everywhere as well. In particular at $s = \bar{s}$. Hence, $P' \mid_{s=\bar{s}}$is well defined, and so either $P' \mid_{s=\bar{s}} = 0$ or $P' \mid_{s=\bar{s}} \neq 0$.

If $P' \mid_{s=\bar{s}} = 0$, then it must be that $s^*(t) \neq \bar{s}$ for any $t \neq \bar{s}$, because for $t \neq \bar{s}$, a small enough deviation from $\bar{s}$ towards t decreases D without increasing P. Thus there is no positive mass of individuals at $\bar{s}$, so it cannot be the norm.

If $P' \mid_{s=\bar{s}} \neq 0$ then either $P' \mid_{s=\bar{s}} > 0$ or $P' \mid_{s=\bar{s}} < 0$. If $P' \mid_{s=\bar{s}} > 0$, then (1) no type with $t < \bar{s}$ will state the norm, as deviating in the left direction from $\bar{s}$ reduces both P and D, and (2) at most one type with $t > \bar{s}$ can have $|D'(\bar{s}; t)| = |P'(\bar{s})|$ when D is strictly concave or strictly convex (i.e., when $\alpha \neq 1$), and so only this one type can have a local min point of L at $\bar{s}$. This means that when $\alpha \neq 1$ there can be no positive mass at $\bar{s}$, which violates the definition of a norm. Now suppose $\alpha = 1$ so that $D = |t - \bar{s}|$. Then each type either speaks her mind or states a statement s such that $|P'(s)| = 1$. Then there can potentially be multiple types choosing the same stance $\bar{s}$ such that $|P'(\bar{s})| = 1$, which implies $\bar{s}$ can be a norm. Suppose this holds and that $\bar{s}$ is the unique norm. Then the fact that no type with $t < \bar{s}$ states $\bar{s}$, together with (i) the uniqueness of the norm $\bar{s}$ and (ii) the fact that a type who does not state an s such that $|P'(s)| = 1$ necessarily speaks her mind, imply a uniform distribution of stances to the left of $\bar{s}$, stemming from the choices of types at this range to speak their minds (this is necessarily so since (a) there must be a finite number of points with $s < \bar{s}$ and $|P'(s)| = 1$ implying that if $s \neq t$ for a positive mass of types with $t < \bar{s}$ then uniqueness of the norm is violated, and (b) all types with $t > \bar{s}$ state $s \geq \bar{s}$ because they either speak their minds or choose the unique norm). The shape of the pressure imposed by the uniform part at $s = [-1, \bar{s}]$ is symmetric around its center, creating the same slope at both edges of this part, $s = -1$ and $s = \bar{s}$. On top of it, there is the pressure stemming from stances $s \geq \bar{s}$. As $\beta > 1$, each of these sources of pressure implies a steeper slope at $s = -1$ than at $s = \bar{s}$, which altogether means that $|P'(-1)| > |P'(\bar{s})| = 1$. This implies that types close to $t = -1$ will gain by deviating to the right from their bliss points, in contradiction to the assumption that they speak their minds The same argument applies when $P' \mid_{s=\bar{s}} < 0$.■

C.2 Alienating societies

C.2.1 Proof of Lemma 1

The minimization problem of the individual is

$$\min_s L(s; t; S) = P(s; S) + |s - t|^{\alpha}. \tag{15}$$

Suppose a single norm exists with a share x stating it. Then

$$P(s) = \begin{cases} K \text{ if } s \neq \bar{s} \\ (1-x)K \text{ if } s = \bar{s} \end{cases}. \tag{16}$$

Therefore $L(s;t;S)$ is increasing in $|s-t|$ except potentially at $s=\bar{s}$, where $P(s) < K$. Thus it is immediate that for each type t, $s^*(t)$ will be either t or $\bar{s}$. Moreover, it is immediate that $s^*(t) = t$ if and only if xK, the difference between $P(t)$ and $P(\bar{s})$, falls below $|t-\bar{s}|^\alpha$, thus follows the lemma.

C.2.2 Proof of Lemma 2

If $y \leq 1 - |\bar{s}|$, the norm is sufficiently centered so that y types on each side follow the norm, which implies $x = y$. When $1 - |\bar{s}| < y \leq 1 + |\bar{s}|$, the norm is sufficiently biased, say to the left, so that there are no longer y types to the left of the norm stating the norm. Then, the total number of individuals declaring the norm is the distance from -1 to $\bar{s}$ on the left and y types on the right. It then follows that the share of norm followers is $x = (y + 1 - |\bar{s}|)/2$. Finally, when $y > 1 + |\bar{s}|$, we get that even the type who is the furthest away from the norm (i.e. at distance $1 + |\bar{s}|$ from it) follows it, implying that all types declare the norm.

C.2.3 Proof of proposition 2

Since Lemma 1 implies that, given a single norm with a share x of followers, $s^*(t)$ is according to (10), a necessary and sufficient condition for this $s^*(t)$ to be the distribution of stances in a single norm equilibrium is that $x(y)$ that is obtained from this distribution of stances in Lemma 2 would equal the value of x that was initially assumed in Lemma 1 for creating this particular $s^*(t)$. This is more conveniently written as a dynamic process, where the requirement is to have $x_{i+1}(y_{i+1}(x_i)) = x_i$. Using (9) and (12) we can write

$$x_{i+1} = f(x_i; K, |\bar{s}|) \equiv \begin{cases} (x_i K)^{1/\alpha} \text{ if } (x_i K)^{1/\alpha} \leq 1 - |\bar{s}| \\ \frac{(x_i K)^{1/\alpha} + 1 - |\bar{s}|}{2} \text{ if } 1 - |\bar{s}| < (x_i K)^{1/\alpha} < 1 + |\bar{s}| \\ 1 \text{ if } (x_i K)^{1/\alpha} \geq 1 + |\bar{s}| \end{cases}. \tag{17}$$

We start by proving parts (1) and (2) of the proposition for the case of $\alpha \geq 1$. If one of the following holds: (1) $\alpha > 1$; (2) $\alpha = 1$, $|\bar{s}| < 1$ and $K \geq 1$; or (3) $\alpha = 1$, $|\bar{s}| = 1$ and $K \geq 2$; then $\lim_{x_i \to +0} f'(x_i; K, |\bar{s}|) \geq 1$, so that $f(x_i; K, |\bar{s}|)$ starts (weakly) above the 45 degree line. In this case, the continuity of $f(x_i; K, |\bar{s}|)$ and the fact that $f(x_i = 1) \leq 1$ imply that $f(x_i; K, |\bar{s}|)$ crosses the 45 degree line at least once in the range $x_i \in (0, 1]$, with the crossing point(s) constituting single norm EQ. Alternatively, if $\alpha = 1$ and either (1) $|\bar{s}| < 1$ and $K < 1$; or (2) $|\bar{s}| = 1$ and $K < 2$; then $f(x_i; K, |\bar{s}|)$ is linear in parts and $\lim_{x_i \to +0} f'(x_i; K, |\bar{s}|) < 1$, so that the first linear part is below the 45 degree line. As the slope of $f(x_i; K, |\bar{s}|)$ only decreases when moving from the first linear

part to the second and from the second to the third, we get that $f(x_i; K, |\bar{s}|)$ is below the 45 degree line in $x_i \in (0,1]$, in which case there is no single norm equilibrium with a strictly positive share x of norm followers. All in all we get that when $\alpha > 1$, $K_{\min}(|\bar{s}|) = 0$; when $\alpha = 1$ and $|\bar{s}| < 1$, $K_{\min}(|\bar{s}|) = 1$; and when $\alpha = 1$ and $|\bar{s}| = 1$, $K_{\min}(|\bar{s}|) = 2$. These values of $K_{\min}(|\bar{s}|)$ are independent of $|\bar{s}|$, which concludes the proof of parts (1) and (2) of the proposition for the case of $\alpha \geq 1$.

We now proceed to proving parts (1) and (2) of the proposition for the case of $\alpha < 1$. To do so, we will now assume that a single norm equilibrium exists at $|\bar{s}|$ and prove the existence of a value $K_{\min}(|\bar{s}|)$ such that the assumption holds if and only if $K \geq K_{\min}(|\bar{s}|)$, and that $K_{\min}(|\bar{s}|)$ is increasing in $|\bar{s}|$.

Looking at the borders between regions in equation (17), we get that if $K \geq (1+|\bar{s}|)^{\alpha}$ then at $x_i = 1$ we are in the third region, implying that $x_{i+1}(x_i) = x_i$ at $x_i = 1$, hence a single norm equilibrium exists (with full compliance to the norm). Otherwise, $(x_i K)^{1/\alpha} \leq K^{1/\alpha} < 1 + |\bar{s}|$, and the third region is irrelevant. Moreover, x_{i+1} in the second region is strictly smaller than 1 and so $x_i = 1$ is not an equilibrium.

Define now

$$\begin{aligned} G(x_i; K, |\bar{s}|) &\equiv x_{i+1}(x_i) - x_i = f(x_i; K, |\bar{s}|) - x_i, \\ &= \begin{cases} (x_i K)^{1/\alpha} - x_i \text{ if } (x_i K)^{1/\alpha} \leq 1 - |\bar{s}| \\ \frac{(x_i K)^{1/\alpha} + 1 - |\bar{s}|}{2} - x_i \text{ if } 1 - |\bar{s}| < (x_i K)^{1/\alpha} < 1 + |\bar{s}| \\ 1 - x_i \text{ if } (x_i K)^{1/\alpha} \geq 1 + |\bar{s}| \end{cases} \end{aligned} \tag{18}$$

which in a single norm equilibrium equals zero for some $x_i \neq 0$. G is continuous in x_i, K and $|\bar{s}|$, with $G(0; K, |\bar{s}|) = 0$ and $G'(0; K, |\bar{s}|) < 0$, and when $K^{1/\alpha} < 1 + |\bar{s}|$ we also get that $G(1; K, |\bar{s}|) < 0$. Differentiation of G with respect to x_i yields

$$G(x_i; K, |\bar{s}|) = \begin{cases} \frac{1}{\alpha} K^{1/\alpha} (x_i)^{1/\alpha - 1} - 1 \text{ if } (x_i K)^{1/\alpha} < 1 - |\bar{s}| \\ \frac{1}{2\alpha} K^{1/\alpha} (x_i)^{1/\alpha - 1} - 1 \text{ if } 1 - |\bar{s}| < (x_i K)^{1/\alpha} < 1 + |\bar{s}| \\ -1 \text{ if } (x_i K)^{1/\alpha} > 1 + |\bar{s}| \end{cases} \tag{19}$$

and

$$G''(x_i; K, |\bar{s}|) = \begin{cases} \frac{1}{\alpha}\left(\frac{1}{\alpha} - 1\right) K^{1/\alpha} (x_i)^{1/\alpha - 2} \text{ if } (x_i K)^{1/\alpha} < 1 - |\bar{s}| \\ \frac{1}{2\alpha}\left(\frac{1}{\alpha} - 1\right) K^{1/\alpha} (x_i)^{1/\alpha - 2} \text{ if } 1 - |\bar{s}| < (x_i K)^{1/\alpha} < 1 + |\bar{s}| \\ 0 \text{ if } (x_i K)^{1/\alpha} > 1 + |\bar{s}| \end{cases} \tag{20}$$

which immediately shows G is strictly convex in the first two regions. It thus follows that when $K^{1/\alpha} < 1 + |\bar{s}|$, G can get a local max only at the border between these two regions, where $x_i = (1 - |\bar{s}|)^{\alpha} / K$. Therefore, when $K^{1/\alpha} < 1 + |\bar{s}|$, there exists a single norm equilibrium if and only if the borderline point falls within the range $[0,1]$

and G at this point is weakly positive.[56] Substituting $x_i = (1-|\bar{s}|)^{\alpha}/K$ in equation (18) yields $G = (1-|\bar{s}|) - (1-|\bar{s}|)^{\alpha}/K$, which equals 0 when $K = (1-|\bar{s}|)^{\alpha-1}$. Substituting this value of K back in x_i we get that $x_i = 1-|\bar{s}|$, thus falls within the range $[0,1]$, and so there exists a single norm equilibrium for $K = (1-|\bar{s}|)^{\alpha-1}$. If K is larger, then the value of x_i at the border between the regions is smaller (hence falls within the range $[0,1]$ too), and the value of G at this point is larger, i.e., positive.

As a result, if we let

$$K_{\min}(|\bar{s}|) \equiv \min\left\{(1-|\bar{s}|)^{\alpha-1}, (1+|\bar{s}|)^{\alpha}\right\}, \tag{21}$$

then for $K < K_{\min}(|\bar{s}|)$ no single norm equilibrium exists, while for any $K \geq K_{\min}(|\bar{s}|)$ there exists a single norm equilibrium at $|\bar{s}|$. It is also worth noting that if $K = K_{\min}(|\bar{s}|)$, the analysis above implies that $\max_{x_i} G(x_i) = 0$ (and reached either at the border between the two regions, if $K_{\min}(|\bar{s}|) = (1-|\bar{s}|)^{\alpha-1}$, or at $x_i = 1$, if $K_{\min}(|\bar{s}|) = (1+|\bar{s}|)^{\alpha}$); while if $K > K_{\min}(|\bar{s}|)$, then $G(x_i) > 0$ either at the borderline point or at $x_i = 1$.

Finally, the fact that $K_{\min}(|\bar{s}|)$ is increasing in $|\bar{s}|$ follows directly from the fact that $(1-|\bar{s}|)^{\alpha-1}$ and $(1+|\bar{s}|)^{\alpha}$ are both increasing in $|\bar{s}|$.■

C.2.4 Proof of proposition 3

We first remind, that in the proposition we treat the unstable steady states (x_{uss}) as ones where if $x_i = x_{uss}$ then $x_{i+1} < x_{uss}$. This includes the cases where $x_{uss} = x_{conv}$. In the proof we do not make this shortcut.

We start with proving all statements of the proposition for the case $\alpha \geq 1$.

When $\alpha > 1$ we get that $\lim_{x_i \to +0} f'(x_i; K, |\bar{s}|) = K/\alpha \lim_{x_i \to +0} (x_i K)^{1/\alpha - 1} = \infty$, implying that there is convergence to the single norm equilibrium (whose existence was proved above – see the proof of Proposition 2 – for every $K > K_{\min}(|\bar{s}|) = 0$) from every $x_i > 0$. It thus follows that in this case $x_{conv}(|\bar{s}|) = 0$ and so (1) $x_{conv}(|\bar{s}|)$ is independent of $|\bar{s}|$ and K; (2) $x_{ss}(|\bar{s}|) > x_{conv}(|\bar{s}|)$; and (3) if $0 \leq x_i \leq x_{conv}(|\bar{s}|)$, then it must be the case that $x_i = 0$ and so $x_{i+1} = x_i = 0$, i.e., there is convergence to a stable steady state where each type speaks her mind ($x_{ss}(|\bar{s}|) = 0$). Moreover, increasing $|\bar{s}|$ has the effect of increasing $x_{ss}(|\bar{s}|)$, as it everywhere weakly decreases x_{i+1} as a function of x_i (by increasing region (2) in equation (17), and since the function f in this region is smaller the larger is $|\bar{s}|$).

When $\alpha = 1$ we have two separate cases to consider. The first one is when $|\bar{s}| < 1$ and $K_{\min}(|\bar{s}|)$ was shown (see the proof of Proposition 2) to equal 1. Here the function f is piecewise linear, where for $K < 1$ it stays below the 45 degree line (see the proof of Proposition 2) and so there is no single norm equilibrium; and for $K > 1$ it stays above the 45 degree line until it reaches 1 and stays there (see equation 17), implying

[56] Note that if the borderline point falls outside the range $[0,1]$, it means that only the first region applies, and then the convexity of G means that $G(1, K, |\bar{s}|) < 0 \Rightarrow G(x_i, K, |\bar{s}|) < 0 \ \ \forall x_i \in]0,1]$, hence no single norm equilibrium exists (we know that $G(1, K, |\bar{s}|) < 0$ because $K^{1/\alpha} < 1 + |\bar{s}|$).

a single norm equilibrium at $x_i = 1$. Thus $x_{ss}(|\bar{s}|) = 1$ (hence independent of $|\bar{s}|$), and there exists a stable steady state if and only if $K > K_{\min}(|\bar{s}|) = 1$ with a share of followers $x_{ss}(|\bar{s}|)$ (if $K = 1$ the function f lies on the 45 line degree in the first region, and so there is a continuum of steady states but none is stable). It follows that if and only if $K > 1$ then there is convergence to $x_{ss}(|\bar{s}|)$ from any $x_i > 0$. Hence, $x_{conv}(|\bar{s}|) = 0$. To complete the proof for this case, note that statement (3) of the proposition holds for $x_{conv}(|\bar{s}|) = 0$ and so $x_{conv}(|\bar{s}|)$ is independent of $|\bar{s}|$ and K, which proves statement (4) too. Finally we should consider the second case, where $\alpha = 1$ and $|\bar{s}| = 1$ (and it was shown in the proof of Proposition 2 that $K_{\min}(|\bar{s}|) = 2$). Here the function f (see equation 17) starts immediately in region (2), and is above the 45 degree line if and only if $K > 2$. The same arguments used in proving the case $|\bar{s}| < 1$ earlier then apply here, with $x_{ss}(|\bar{s}|) = 1$ and $x_{conv}(|\bar{s}|) = 0$.

The proof of the proposition for the case $\alpha < 1$ builds on a few preliminary results and auxiliary lemmas, which will be presented first.

Note first that Lemmas 1 and 2 show that alienation recreates alienation. Hence, the full dynamics can be described by the dynamics of x, the share of norm followers, as given in equation (18). Following equation (20), it is straightforward to see that $x_{i+1} = f(x_i; K, |\bar{s}|)$ is convex within each of the first two regions and has a kink at the border between the regions. Together with $G'(0; K, |\bar{s}|) < 0$ (see equation (19)), this means we can define the following values of x_{i+1} (see Figure 8) that exhaust the possible fix points, and which will be used throughout the upcoming lemmas.

$$\begin{aligned}
\hat{x} &\equiv \{x_i : x_{i+1} = x_i \text{ and } x_i \text{ is in the first region}\} \qquad (22)\\
\check{x} &\equiv \{x_i : x_{i+1} = x_i \text{ and } x_i \text{ is in the second region and } G' > 0\}\\
\tilde{x} &\equiv \{x_i : x_{i+1} = x_i \text{ and } x_i \text{ is in the second region and } G' < 0\}\\
\ddot{x} &\equiv \left\{x_i : (x_i K)^{1/\alpha} = 1 - |\bar{s}| \right\} \text{ (i.e., at the border between regions (1) and (2))}\\
\dot{x} &\equiv \left\{x_i : (x_i K)^{1/\alpha} = 1 - |\bar{s}| \text{ and } G(x_i) = 0 \text{ and } G_2'(x_i) < 0\right\}\\
x_{end} &\equiv \{x_i : x_{i+1} = x_i = 1\} \text{ (i.e., at the endpoint)}
\end{aligned}$$

Note that when $G(\ddot{x}) = 0$ then either $G_2'(x_i) < 0$, in which case $\ddot{x} = \dot{x}$, or $G_2'(x_i) > 0$.

Lemma 6 *Consider a given x_i. Then $G'(x_i : x_i < \ddot{x}) > G'(x_i : x_i > \ddot{x})$.*
Proof. *Let G_1, G_2 and G_3 denote the values of G in regions (1), (2) and (3) respectively. When $x_i < \ddot{x}$, G_1 applies, and when $x_i > \ddot{x}$, G_2 applies. Then for a given x_i, $G_1' = \frac{1}{\alpha} K^{1/\alpha} (x_i)^{1/\alpha - 1} - 1 > \frac{1}{2\alpha} K^{1/\alpha} (x_i)^{1/\alpha - 1} - 1 = G_2'$.* ■

Lemma 7 *G' is weakly falling in $|\bar{s}|$ for any $x_i < (1 + |\bar{s}|)^{\alpha} / K$.*
Proof. *When $x_i < (1 + |\bar{s}|)^{\alpha} / K$ we are in region (1) or region (2) of equation (19). Here, $\frac{dG_1'}{d|\bar{s}|} = \frac{dG_2'}{d|\bar{s}|} = 0$. Moreover, $\ddot{x} = (1 - |\bar{s}|)^{\alpha} / K$ decreases in $|\bar{s}|$. This implies that*

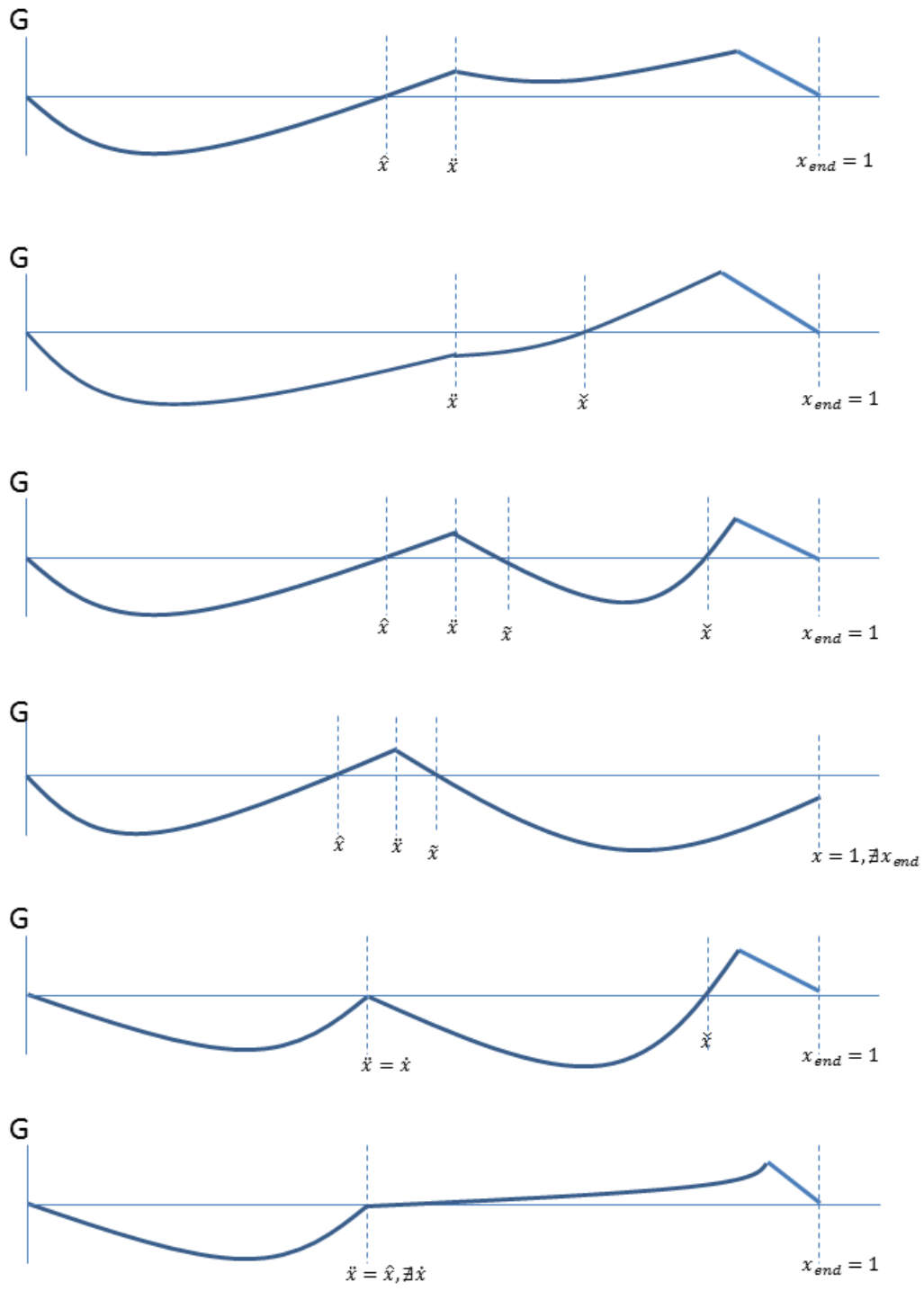

Figure 8: Some variations of the G function of equation (18), depicting the potential fix points defined in equation (22). Note that these variations of G are not exhaustive but are only meant to complement the proof.

if $|\bar{s}|$ increases, region (2) expands at the expense of region (1). Then, by Lemma 6, we get that G' is weakly falling in $|\bar{s}|$. ■

Lemma 8 *1) If $\hat{x}$ exists then it is independent of $|\bar{s}|$. 2) If $\check{x}$ exists then it is weakly increasing in $|\bar{s}|$. 3) If $\tilde{x}$ exists it is weakly decreasing in $|\bar{s}|$.*
Proof. *1) By definition $\hat{x}$ is in region 1. Hence G_1 applies. Since G_1 is independent of $|\bar{s}|$ so must $\hat{x}$ be. 2) By definition $\check{x}$ is in region 2. Lemma 7 together with $G(0) = 0$ imply that G is weakly falling in $|\bar{s}|$ in region 1 and 2. Combined with the fact that $G'(\check{x}) > 0$ (by definition) this implies $\check{x}$ (if it exists) is weakly increasing in $|\bar{s}|$. 3) Same logic as part 2 but now with $G'(\tilde{x}) < 0$.* ■

Lemma 9 *If $\exists\hat{x}$ for some $|\bar{s}|$ then $\exists\hat{x}$ for any $|\bar{s}'| < |\bar{s}|$.*
Proof. *G_1 is independent of $|\bar{s}|$. Then the fact that $|\bar{s}'| < |\bar{s}|$ implies that region (1) is broader under $|\bar{s}'|$, so if $\exists\hat{x}$ for some $|\bar{s}|$ then $\exists\hat{x}$ for any $|\bar{s}'| < |\bar{s}|$.* ■

Lemma 10 *Suppose $K > K_{\min}$. Let $x_{conv} \equiv \{x_i : x_i = \min\{\hat{x}, \check{x}\}\}$ (when $\hat{x}$ or $\check{x}$ or both exist). Then:*

1. *If $x_i > x_{conv}(|\bar{s}|)$ there is convergence to a stable single norm steady state where a share $x_{ss}(|\bar{s}|) > x_{conv}(|\bar{s}|)$ of the population state $\bar{s}$.*

2. *Otherwise, provided that $\nexists\dot{x}$, if $0 \leq x_i < x_{conv}(|\bar{s}|)$, there is convergence to a stable steady state where each type speaks her mind ($x_{ss}(|\bar{s}|) = 0$).*

3. *Furthermore, if $\exists\dot{x}$, then when $0 < x_i < \dot{x}$ there is convergence to a stable steady state where each type speaks her mind ($x_{ss}(|\bar{s}|) = 0$), and when $\dot{x} \leq x_i < x_{conv}$ there is convergence to an unstable single norm steady state where a share $\dot{x}$ state the norm.*[57]

4. *x_{conv} increases in $|\bar{s}|$.*

5. *x_{conv} decreases in K.*

Proof. *We start with statement 2) $G'(\hat{x}) > 0$ since $G_1(0) = 0$, $G_1'(0) < 0$ and G_1 is convex. $G'(\check{x}) > 0$ by definition. This implies $\hat{x}$ and $\check{x}$ are unstable steady states. Furthermore, they are the only unstable states.*[58] *Hence, if $\hat{x}$ exists, it must be the smallest strictly positive steady state, and so $G_1(0) = 0$ and $G_1'(0) < 0$ imply that $\forall x_i < \hat{x} = x_{conv}$ we have $G(x_i) < 0$, i.e., $x_{i+1} < x_i$. Otherwise there is no steady*

[57] In line with our general treatment of unstable steady states as converging to less conformity, statement (3) in the proposition treats this special case as one where x_i, upon reaching $\dot{x}$, only passes through it and continues to $x_{ss} = 0$.

[58] To see this note that $\tilde{x}$ must be stable by $G'(\tilde{x}) < 0$. Furthermore, recall that $\nexists\dot{x}$. Hence, the only way for $\ddot{x}$ to be a steady state is if $G(\ddot{x}) = 0$ and $G'(\ddot{x}) > 0$, which implies $\ddot{x} = \hat{x}$ (see above). Finally, if x_{end} exists in region 3 it must be stable since $G_3' < 0$ and if x_{end} exists in region 1 or 2 then it must be that either $x_{end} = \check{x}$ or $x_{end} = \hat{x}$.

state in the first region, in which case $\check{x}$ must be the smallest strictly positive steady state. Then again $G_1(0) = 0$ and $G_1'(0) < 0$ imply that $x_{i+1} < x_i \ \ \forall x_i < x_{conv}$. Thus, the instability of x_{conv} implies that $x_{ss}(|\bar{s}|) = 0$.

1) In the proof of Proposition 2 we showed that $G > 0$ for some x_i iff $K > K_{\min}$. This implies $\hat{x}$ or $\check{x}$ or both exist. Since $G' > 0$ at both, this implies $x_{i+1} > x_i$ in a neighborhood of $x_i > x_{conv}$, which implies convergence to a stable steady state.

3) When $\exists \dot{x}$, we know by convexity of G_1 (and since the definition of $\hat{x}$ requires that $G' > 0$ at $\hat{x}$) that $\hat{x}$ does not exist. Hence, the only possible fix points are $\dot{x}, \check{x}$ and x_{end}. Note that by the definition of $\dot{x}$ it must be stable in a neighborhood above $\dot{x}$. By convexity of G_1, $\dot{x}$ must be unstable from below. Since there are no other fix points below $\dot{x}$, $x_i < \dot{x}$ implies convergence to $x_{ss} = 0$. This concludes the first subsentence. Furthermore, by instability of $\check{x}$ and stability of $\dot{x}$ from above we know that if $x_i \in]\dot{x}, x_{conv}[$, then there will be convergence to $\dot{x}$ which implies the second subsentence.

4) $x_{conv} \equiv \hat{x}$ whenever $\exists \hat{x}$. From Lemma 8 we know that $\hat{x}$ is independent of $|\bar{s}|$ and from Lemma 9 we know it exists iff $|\bar{s}|$ is sufficiently small. Hence, as $|\bar{s}|$ is increased, x_{conv} is either constant, or it makes a discrete jump to equal $\check{x}$ (which we know exists since $K > K_{\min}$ while in this scenario $\hat{x}$ seizes to exist). Furthermore, by Lemma 8 we know $\check{x}$ is increasing in $|\bar{s}|$. Put together, this implies that x_{conv} is either constant or increasing in $|\bar{s}|$.

5) By definition of $\hat{x}$ we get $\hat{x} = K^{1/(\alpha-1)}$, which decreases in K. By definition of $\check{x}$ and using equation (18) we get an implicit expression $H = (\check{x}K)^{1/\alpha} + 1 - |\bar{s}| - 2\check{x} = 0$ defining $\check{x}$. Using the implicit function theorem we get $d\check{x}/dK = -(\check{x})^{1/\alpha} K^{1/\alpha - 1}/\alpha / \left(K^{1/\alpha}(\check{x})^{1/\alpha-1}/\alpha - 2\right) < 0 \Leftrightarrow K^{1/\alpha}(\check{x})^{1/\alpha-1} > 2\alpha$. From equation (19) this condition corresponds to the condition for $G_2' > 0$, which holds by the definition of $\check{x}$. Hence x_{conv} is locally decreasing in K. Note now that, by equation (18), G_1 and G_2 are increasing in K. Hence, as K increases, we cannot switch from $x_{conv} = \hat{x}$ to $x_{conv} = \check{x}$. This implies that x_{conv} is decreasing in K also globally. ■

Lemma 11 *Suppose $K > K_{\min}$. Then there exists a stable steady state with a single norm at $x_{ss} = \tilde{x}$ or at $x_{ss} = 1$ or at both. Moreover, x_{ss} is weakly decreasing in $|\bar{s}|$.*

Proof. *When $K > K_{\min}$, a stable steady state must exist (see the proof of Proposition 2). All the steady states except for $\tilde{x}$ and 1 must be unstable since they all imply $G' > 0$ on at least one side of the steady state. Hence, when $K > K_{\min}$ there exists a stable steady state at $x_{ss} = \tilde{x}$ or at $x_{ss} = 1$ or at both, and since $x_{ss} \neq 0$, the steady state contains a single norm. If $\tilde{x}$ exists, we know from Lemma 8 that $\tilde{x}$ is strictly decreasing in $|\bar{s}|$. As for $x_{ss} = x_{end} = 1$, it is constant in $|\bar{s}|$, and it is stable if and only if region (3) is reached for some $x_i < 1$, i.e., iff $(1 \cdot K)^{1/\alpha} > 1 + |\bar{s}|$ (this inequality is obtained by plugging in $x_i = 1$ in the border between regions (2) and (3) in equation (18)). Therefore, as $|\bar{s}|$ increases, the steady states can only decrease from $x_{ss} = 1$ to $x_{ss} = \tilde{x}$ (but not the other way around).* ■

Proof of proposition 3

Part 1: The 'if' part follows from Lemma 11. As for the 'only if' part, we showed in the proof of Proposition 2 that the function G is strictly positive at some point iff $K > K_{\min}$. Hence, if $K \leq K_{\min}$, then $\forall x_i$ we have $x_{i+1} \leq x_i$, which means that there can be no convergence from the left to any steady state, implying that a stable steady state with a single norm cannot exist.

Part 2: Follows from Lemma 11.

Part 3: Follows from Lemma 10.

Part 4: Follows from Lemma 10.■

C.3 Inverting societies

Let

$$s_l \equiv \bar{s} + 1 \text{ and}$$

$$\sigma \equiv s - \bar{s}.$$

These notations will be useful for proofs that deal with the case in which $\bar{s} < 0$ and $y > \bar{s} + 1$, where the distribution of stances is asymmetric around $\bar{s}$, and s_l then denotes the size of the uniform part to the left of $\bar{s}$, which equals the distance of $\bar{s}$ from the left corner of the types distribution, -1.

C.3.1 Proof of Lemma 3

When D is a step function taking the value of 0 or 1 and $P(s)$ is monotonically increasing in $|s - \bar{s}|$, we immediately have

$$s^*(t) = \begin{cases} \bar{s} \text{ if } 1 + P(\bar{s}) \leq P(t) \\ t \text{ if } 1 + P(\bar{s}) > P(t) \end{cases}. \tag{23}$$

Since $P(t)$ is increasing in $|t - \bar{s}|$, types sufficiently far from the norm will state the norm and types sufficiently close to the norm will state their type.

C.3.2 Proof of Lemma 4

If $[\bar{s} - y, \bar{s} + y] \cap [-1, 1] = [\bar{s} - y, \bar{s} + y]$, the distribution of stances is composed of a mass of individuals at $\bar{s}$ and a uniform part that is symmetric around $\bar{s}$. The pressure that results from each of the two parts of this distribution of stances increases in the distance from $\bar{s}$ (see Lemma 5 regarding the contribution from the uniform part), and so the lemma holds. Otherwise, assume without loss of generality that $\bar{s} < 0$ and that all types at $[-1, \bar{s} + y]$ speak their minds, with $y > \bar{s} + 1$. The aggregate $P(s)$ that results from this distribution of stances can be written as

$$P(s) = \begin{cases} Kx|s - \bar{s}|^{\beta} + K\frac{1}{2}\frac{(s+1)^{\beta+1} + (\bar{s}+y-s)^{\beta+1}}{\beta+1} \text{ if } s \leq \bar{s} + y \\ Kx|s - \bar{s}|^{\beta} + K\frac{1}{2}\frac{(s+1)^{\beta+1} - (s-\bar{s}-y)^{\beta+1}}{\beta+1} \text{ if } s > \bar{s} + y \end{cases} \tag{24}$$

with

$$x = \left(1 - \frac{y}{2} - \frac{\bar{s}+1}{2}\right).$$

From the following expression of $P'(s)$

$$P'(s) = \begin{cases} -K\left(1-\frac{y}{2}-\frac{\bar{s}+1}{2}\right)\beta\left(\bar{s}-s\right)^{\beta-1} + K\frac{1}{2}\left(s+1\right)^{\beta} - K\frac{1}{2}\left(\bar{s}+y-s\right)^{\beta} & \text{if } s < \bar{s} \\ K\left(1-\frac{y}{2}-\frac{\bar{s}+1}{2}\right)\beta\left(s-\bar{s}\right)^{\beta-1} + K\frac{1}{2}\left(s+1\right)^{\beta} - K\frac{1}{2}\left(\bar{s}+y-s\right)^{\beta} & \text{if } \bar{s} < s \leq \bar{s}+y \\ K\left(1-\frac{y}{2}-\frac{\bar{s}+1}{2}\right)\beta\left(s-\bar{s}\right)^{\beta-1} + K\frac{1}{2}\left(s+1\right)^{\beta} - K\frac{1}{2}\left(s-\bar{s}-y\right)^{\beta} & \text{if } s > \bar{s}+y \end{cases} \tag{25}$$

it is clear that $P(s)$ is decreasing in s for $s < \bar{s}$ (recall that $y > \bar{s}+1$) and is increasing in s for $s > \bar{s}+y$. Moreover, when $\frac{-1+\bar{s}+y}{2} < s \leq \bar{s}+y$ (i.e., s in the right half of the uniform part), we get that $(s+1) > (\bar{s}+y-s)$, hence $P'(s)$ is positive too (this comes from the fact that the part of $P(s)$ that originates in the uniform part is increasing in the distance from $\frac{-1+\bar{s}+y}{2}$, the center of this part). Therefore, the global min can only be found at $s \in \left[\bar{s}, \frac{-1+\bar{s}+y}{2}\right]$. In this range we have

$$P'(s) = K\left(1-\frac{y}{2}-\frac{\bar{s}+1}{2}\right)\beta\left(s-\bar{s}\right)^{\beta-1} + K\frac{1}{2}\left(s+1\right)^{\beta} - K\frac{1}{2}\left(\bar{s}+y-s\right)^{\beta}.$$

Note first that (i) if $y = \bar{s}+1$, the distribution of stances is symmetric around $\bar{s}$, and so $P'(s) \geq 0$ at the range $s \in \left[\bar{s}, \frac{-1+\bar{s}+y}{2}\right]$; and (ii) if $y = 1-\bar{s}$ (this is the distance from $\bar{s}$ to the furthest edge), then $P'(s) < 0$ at the range $s \in \left[\bar{s}, \frac{-1+\bar{s}+y}{2}\right]$, since Lemma 5 implies that $P(s)$ is increasing in the distance from $0 > \frac{-1+\bar{s}+y}{2}$. Differentiating with respect to y we get

$$\frac{dP'(s)}{dy} = \frac{1}{2}K\left[-\beta\left(s-\bar{s}\right)^{\beta-1} - \beta\left(\bar{s}+y-s\right)^{\beta-1}\right] < 0 \tag{26}$$

This inequality, together with i) and ii), then implies that $\exists y \in \left]\bar{s}+1, 1-\bar{s}\right[$, denoted by $y_{\max}(\bar{s})$, such that $P'(s) \geq 0$ at the whole range $s \in \left[\bar{s}, \frac{-1+\bar{s}+y}{2}\right]$ if and only if $y \leq y_{\max}(\bar{s})$.[59] We will now show that $y_{\max}(\bar{s}) \geq 1$, by showing that for $y = 1$ and every given $\bar{s}$, $P'(s) \geq 0$ at the whole range $s \in \left[\bar{s}, \frac{-1+\bar{s}+y}{2}\right]$.

Rewriting the expression for $P'(s)$ we get

$$P'(s) = \frac{1}{2}K\left[\left(2-y-s_l\right)\beta\sigma^{\beta-1} + \left(s_l+\sigma\right)^{\beta} - \left(y-\sigma\right)^{\beta}\right]. \tag{27}$$

Differentiating with respect to s_l we get

$$\frac{dP'(s)}{ds_l} = \frac{1}{2}K\left[-\beta\sigma^{\beta-1} + \beta\left(s_l+\sigma\right)^{\beta-1}\right] \leq 0 \tag{28}$$

[59] This already takes into account the fact that the range $\left[\bar{s}, \frac{-1+\bar{s}+y}{2}\right]$ is itself increasing in y.

This inequality suggests that $P'(s)$ is minimal when s_l is maximal (i.e., equals $1-\varepsilon$, where $\bar{s}=-\varepsilon\to 0$). Note that in this case $\sigma\to 0$, as the range of s shrinks to be $s\in\left[-\varepsilon,\frac{-\varepsilon}{2}\right]$. Plugging $s=-\lambda\varepsilon$ into (27), and letting $\lambda\in[0.5,1]$, we then have

$$\begin{aligned}P'(s)&=\frac{\varepsilon}{2}\beta\left(-\lambda\varepsilon+\varepsilon\right)^{\beta-1}+\frac{1}{2}\left(-\lambda\varepsilon+1\right)^{\beta}-\frac{1}{2}\left(-\varepsilon+1+\lambda\varepsilon\right)^{\beta}\\&=\frac{\varepsilon^{\beta}}{2}\beta\left[\left(1-\lambda\right)\right]^{\beta-1}+\frac{1}{2}\left(1-\lambda\varepsilon\right)^{\beta}-\frac{1}{2}\left[1-\left(1-\lambda\right)\varepsilon\right]^{\beta},\end{aligned}$$

we get[60]

$$P'(s)=\frac{\varepsilon^{\beta}}{2}\beta\left[\left(1-\lambda\right)\right]^{\beta-1}+\frac{1}{2}\left[\beta\left(1-2\lambda\right)\varepsilon+O(\varepsilon^{2})\right]$$

and so, if $\beta<1$

$$\lim_{\varepsilon\to 0}P'(s)=\lim_{\varepsilon\to 0}\frac{\varepsilon^{\beta}}{2}\beta\left[\left(1-\lambda\right)\right]^{\beta-1}=0^{+}$$

and if $\beta=1$

$$\lim_{\varepsilon\to 0}P'(s)=\frac{\varepsilon}{2}\left[1+1-2\lambda\right]=0^{+}.$$

This means that even for the maximal s_l, $P'(s)$ is positive everywhere when $y=1$, implying that $y_{\max}(\bar{s})\geq 1$.

C.3.3 Proof of Proposition 4

The proof of the proposition builds on a few auxiliary lemmas that are outlined first. The actual proof of the proposition follows after the lemmas.

Lemma 12 *If $\beta=1$ then $y_{\max}(\bar{s})=1\ \ \forall\bar{s}$.*
Proof. *Lemma 4 implies that $y_{\max}(\bar{s})\geq 1$. Plugging in $\beta=1$ and letting $s\to^{+}\bar{s}$ in equation 25 yields $P'(s)=K(1-y)$. This expression is negative for $y>1$, which, by the definition of $y_{\max}(\bar{s})$ implies that $y_{\max}(\bar{s})\leq 1$. Thus $y_{\max}(\bar{s})=1\ \ \forall\bar{s}$.* ■

Lemma 13 *Suppose that $\beta<1$ and $s_l\in[0,1]$. Then $(1-s_l)\beta-2+(s_l+1)^{\beta}<0$.*
Proof. *$(1-s_l)\beta\geq 0$ and $2-(s_l+1)^{\beta}\geq 0$. However, we have $(s_l+1)^{\beta}<s_l+1=2-(1-s_l)<2-(1-s_l)\beta$, and so $(1-s_l)\beta-\left[2-(s_l+1)^{\beta}\right]<0$.* ■

Lemma 14 *Suppose $\beta\leq 1$. Let $\bar{s}\leq 0$ and $y\leq y_{\max}(\bar{s})$, and suppose that all types $t\in[\bar{s}-y,\bar{s}+y]\cap[-1,1]$ speak their minds and the rest state $\bar{s}$. If type $t=\bar{s}+y$ is indifferent between the two corner solutions $s^{*}(t)=\bar{s}$ and $s^{*}(t)=t$, then for any type the best response is*

$$s^{*}(t)=\begin{cases}\bar{s} & \text{if } |t-\bar{s}|>y\\ t & \text{otherwise}\end{cases}.$$

[60] In the following expression, $O(\varepsilon^2)$ is the standard mathematical notation for an element in the order of ε^2.

Proof. *For types with $t > \bar{s}$ the result follows from Lemmas 3 and 4. As for types $t < \bar{s}$, if $[\bar{s}-y, \bar{s}+y] \cap [-1,1] = [\bar{s}-y, \bar{s}+y]$ then the distribution of stances is symmetric around $\bar{s}$ and the result follows from P then being symmetric and monotonically increasing in $|s-\bar{s}|$. Otherwise, by construction all types at $[-1, \bar{s}+y]$ speak their minds, where $y > \bar{s}+1$. We need to show that indeed all types with $t < \bar{s}$ have strict preference for the solution $s^*(t) = t$. Since we know from Lemma 4 that P is strictly increasing in the distance from $\bar{s}$ while D is fixed, it is sufficient to show that $s^*(t) = t$ for the type $t = -1$. Looking at $t = -1$, the fact that P gets its global min point at $\bar{s}$ and equation (23) imply that it is sufficient to show that $1 + P(\bar{s}) - P(-1) \geq 0$. Furthermore, note that the indifference of type $t = \bar{s} + y$ implies that $1 + P(\bar{s}) - P(\bar{s}+y) = 0$. Therefore, it is sufficient to show that $P(\bar{s}+y) \geq P(-1)$:*

$$P(\bar{s}+y) = Kxy^{\beta} + K\frac{1}{2}\frac{(\bar{s}+y+1)^{\beta+1}}{\beta+1},$$
$$P(-1) = Kx\,|-1-\bar{s}|^{\beta} + K\frac{1}{2}\frac{(\bar{s}+y+1)^{\beta+1}}{\beta+1},$$

and so $P(\bar{s}+y) \geq P(-1)$ if and only if $y \geq \bar{s}+1$, which holds by assumption. ■

Lemma 15 *Let $\bar{s} \in [-1,1]$ and let D be given by (14), and suppose that $\beta \leq 1$. For every $y \leq y_{\max}(\bar{s})$, let $S(y)$, the distribution of stances in society, be such that all types $t \in [\bar{s}-y, \bar{s}+y] \cap [-1,1]$ speak their minds while the rest choose $\bar{s}$. Denote by $K(y)$ the value of K that, given the pressure function $P(s)$ that results from $S(y)$, implies indeed $s^*(t) = \bar{s}$ for all types with $|t-\bar{s}| > y$ and $s^*(t) = t$ for all types with $|t-\bar{s}| \leq y$. Then, when $\beta < 1$, $K(y)$ has either a U-shape or a W-shape, and when $\beta = 1$, $K(y)$ is monotonically decreasing.*

Proof. *Without loss of generality, let $\bar{s} \leq 0$. The given distribution of stances and the fact that $y \leq y_{\max}(\bar{s})$ imply by Lemma 4 that P is increasing in $|\sigma|$ (recall $\sigma \equiv s - \bar{s}$). Moreover, from Lemma 3 we know that*

$$s^*(t) = \begin{cases} \bar{s} \text{ if } 1 + P(\bar{s}) \leq P(t) \\ t \text{ if } 1 + P(\bar{s}) > P(t) \end{cases}$$

which implies types sufficiently far from the norm will state the norm and types sufficiently close to the norm will state their type. We are looking for the value of K for which the type who is indifferent between the two options is at distance y from $\bar{s}$. I.e., $1 + P(\bar{s}) = P(\bar{s}+y)$. Lemma 14 implies that this distance y applies to both sides. However, as y grows from 0, we move from a region where the uniform part is symmetric around $\bar{s}$ (when $y \leq s_l$) to a region where it is asymmetric (when $y \in [s_l, 2-s_l]$). Therefore the analysis will be first performed separately for each region, and then the two analyses will be combined.

Region (1): $y \leq s_l$

In this region the uniform part of S is symmetric around the norm and so the share of individuals stating the norm is $x = 1 - y$ and $P(\sigma)$ is given by:

$$P(\sigma) = \begin{cases} Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(|\sigma|+y)^{\beta+1}+(y-|\sigma|)^{\beta+1}}{\beta+1} & \text{if } |\sigma| \leq y \\ Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(|\sigma|+y)^{\beta+1}-(|\sigma|-y)^{\beta+1}}{\beta+1} & \text{if } |\sigma| > y \end{cases} \tag{29}$$

The type who is indifferent between the two options is at distance y from $\bar{s}$, i.e., $1 + P(0) = P(y)$, if

$$\begin{aligned} 1/K + \frac{1}{2}\frac{2y^{\beta+1}}{\beta+1} &= (1-y)y^{\beta} + \frac{1}{2}\frac{(2y)^{\beta+1}}{\beta+1} \\ &\Rightarrow 1/K = (1-y)y^{\beta} + \left(2^{\beta}-1\right)\frac{y^{\beta+1}}{\beta+1}. \end{aligned} \tag{30}$$

Region (2): $y \in [s_l, 2 - s_l]$

In this region the uniform part of S is asymmetric around the norm, and the share of individuals stating the norm is $x = \left(1 - \frac{y}{2} - \frac{s_l}{2}\right)$. Rewriting (24) we get that $P(\sigma)$ is given by:

$$P(\sigma) = \begin{cases} Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(s_l+\sigma)^{\beta+1}+(y-\sigma)^{\beta+1}}{\beta+1} & \text{if } \sigma \leq y \\ Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(s_l+\sigma)^{\beta+1}-(\sigma-y)^{\beta+1}}{\beta+1} & \text{if } \sigma \geq y \end{cases}.$$

The type who is indifferent between the two options is at distance y from $\bar{s}$, i.e., $1 + P(0) = P(y)$, if

$$\begin{aligned} 1/K + \frac{1}{2}\frac{(s_l)^{\beta+1} + (y)^{\beta+1}}{\beta+1} &= \left(1 - \frac{y}{2} - \frac{s_l}{2}\right)y^{\beta} + \frac{1}{2}\frac{(s_l+y)^{\beta+1}}{\beta+1} \Rightarrow \\ 1/K &= \left(1 - \frac{y}{2} - \frac{s_l}{2}\right)y^{\beta} + \frac{1}{2}\frac{(s_l+y)^{\beta+1} - (s_l)^{\beta+1} - y^{\beta+1}}{\beta+1}. \end{aligned} \tag{31}$$

Joining the two regions:

Following equations 30 and 31, we can get the following expression for $\frac{1}{K}$ as a function of y.

$$\frac{1}{K}(y) = \begin{cases} (1-y)y^{\beta} + \left(2^{\beta}-1\right)\frac{y^{\beta+1}}{\beta+1} & \text{if } y \leq s_l \\ \left(1 - \frac{y}{2} - \frac{s_l}{2}\right)y^{\beta} + \frac{1}{2}\frac{(s_l+y)^{\beta+1}-(s_l)^{\beta+1}-y^{\beta+1}}{\beta+1} & \text{if } y \in [s_l, 2-s_l] \end{cases} \tag{32}$$

Differentiating in both regions yields

$$\frac{d(1/K)}{dy} = \begin{cases} (1-y)y^{\beta-1}\beta - y^{\beta}\left(2-2^{\beta}\right) & \text{if } y \leq s_l \\ \left(1 - \frac{y}{2} - \frac{s_l}{2}\right)\beta y^{\beta-1} - y^{\beta} + \frac{1}{2}(s_l+y)^{\beta} & \text{if } y \in [s_l, 2-s_l] \end{cases}. \tag{33}$$

When $\beta = 1$ we get that $\frac{d(1/K)}{dy} = 1 - y$ in both regions, hence $1/K$ is a strictly increasing function of y in the range $[0,1]$ (and $K(y)$ is strictly decreasing in $y \in [0,1]$). Since in this case $y_{\max} = 1$ (see Lemma 12), we get that the lemma holds for $\beta = 1$. We continue now with the case of $\beta < 1$. Differentiating once more

$$\frac{d^2(1/K)}{dy^2} = \begin{cases} -y^{\beta}\beta + (1-y)\, y^{\beta-2}(\beta-1)\beta - \beta y^{\beta-1}\left(2-2^{\beta}\right) < 0 \text{ if } y \leq s_l \\ \left(1-\frac{y}{2}-\frac{s_l}{2}\right)\beta(\beta-1)\, y^{\beta-2} - \frac{3}{2}\beta y^{\beta-1} + \frac{1}{2}\beta(s_l+y)^{\beta-1} < 0 \text{ if } y \in [s_l, 2-s_l] \end{cases} \quad (34)$$

so that $1/K$ is concave in y in both regions. Moreover, it is easy to verify that $\frac{1}{K}(y)$ is continuous at $y = s_l$, the border between the two regions. If $\bar{s} = 0$ ($s_l = 1$), then only the first region applies. It is easy to verify that in the first region we get the following

$$\begin{cases} \frac{d(1/K)}{dy} > 0 \text{ as } y \to 0 \\ \frac{d(1/K)}{dy} < 0 \text{ as } y \to 1 \end{cases},$$

and so in this case $\frac{1}{K}(y)$ is hill-shaped. Otherwise $\bar{s} < 0$ ($s_l < 1$). For the applicability of $\frac{1}{K}(y)$ in this lemma we require that $y \leq y_{\max}(\bar{s})$. When $\bar{s} < 0$ we still have $\frac{d(1/K)}{dy} > 0$ as $y \to 0$, but $s_l < 1 \leq \bar{y} \equiv \min\{y_{\max}(\bar{s}), 2 - s_l\}$ (recall that from Lemma 4 we know that $y_{\max}(\bar{s}) \geq 1$), and so region 2 applies to large enough values of y. Moreover, $\frac{d^2(1/K)}{dy^2} < 0$ implies that $\frac{d(1/K)}{dy}$ is strictly decreasing in y. Hence, $\bar{y} \geq 1$ implies that $\frac{d(1/K)}{dy}|_{y=\bar{y}} \leq \frac{d(1/K)}{dy}|_{y=1} = \frac{1}{2}\left[(1-s_l)\beta - 2 + (s_l+1)^{\beta}\right]$, which by Lemma 13 is strictly negative. Hence we know that $\frac{1}{K}(y)$ has a positive slope at $y \to 0$ and a negative slope at $y = \min\{y_{\max}(\bar{s}), 2 - s_l\}$, and in between it is concave in each of the regions. It thus follows that $\frac{1}{K}(y)$ has at least one and at most two max points and that these max points are internal, i.e. $\frac{1}{K}(y)$ is either hill-shaped or M-shaped, and so $K(y)$ is either U-shaped or W-shaped. ■

Lemma 16 *Let D be given by (14) and let $\beta < 1$. Suppose there exists a value of K such that a single norm equilibrium at $\bar{s} \in [-1,1]$, where all types $t \in [\bar{s} - y, \bar{s} + y] \cap [-1,1]$ speak their minds while the rest choose $\bar{s}$, exists for some $y > y_{\max}(\bar{s})$. Then $K \geq K_{\min}(|\bar{s}|)$.*

Proof. *Without loss of generality, let $\bar{s} \leq 0$. Since the existence of the equilibrium that is described in the lemma requires that $t = \bar{s} + y$ will be indifferent between speaking her mind and choosing $\bar{s}$, and since $y > y_{\max}(\bar{s}) \geq 1 \geq s_l$, the value of K that may allow such an equilibrium (if it indeed exists) is given by equation (31), with first and second derivatives as in the second lines of equations (33) and (34) respectively. Then, the fact that $\frac{d^2(1/K)}{dy^2} < 0$ implies that the value of $\frac{d(1/K)}{dy}$ at any $y > y_{\max}(\bar{s})$ is strictly smaller than $\frac{d(1/K)}{dy}|_{y=1} = \frac{1}{2}\left[(1-s_l)\beta - 2 + (s_l+1)^{\beta}\right]$, which*

by Lemma 13 is negative. Hence, $\frac{1}{K}(y)$ is decreasing when $y > y_{\max}(\bar{s})$, implying that for any $y > y_{\max}(\bar{s})$, an equilibrium as described in the lemma requires $K(y) > K(y_{\max}(\bar{s})) \geq K_{\min}(|\bar{s}|)$. ■

Lemma 17 *Let D be given by (14) and suppose that $\beta \leq 1$. Then the only possible distribution of stances in a single norm equilibrium at $\bar{s} \in [-1,1]$ is one where all types $t \in [\bar{s}-y, \bar{s}+y] \cap [-1,1]$ for some $y > 0$ speak their minds while the rest choose $\bar{s}$.*

Proof. *First note that if D is a step function as in (14), then for any $t \in [-1,1]$, either $s^*(t) = t$ or $s^*(t) \in \arg\min(P)$. Then, the existence of a single norm equilibrium at $\bar{s}$ implies that (i) $\bar{s} \in \arg\min(P)$ and (ii) $s^*(t) = t$ for every t for whom $s^*(t) \neq \bar{s}$. Together with the uniform distribution of types, this implies that the distribution of stances can contain only uniform parts apart from the peak at $\bar{s}$.*

Moreover, the continuity of $P(s)$ implies that for types sufficiently close to $\bar{s}$, $1 + P(\bar{s}) > P(t)$ (since then $P(t) \to P(\bar{s})$), and so the distribution of stances must necessarily contain a uniform part that is attached to $\bar{s}$. We will now show that there can be no other uniform parts in the distribution of stances. Without loss of generality, let $\bar{s} \leq 0$, and suppose that there exist (one or more) uniform parts that are detached from $\bar{s}$. Consider the rightmost uniform part. Since P is continuous, at the left edge of this specific uniform part there must be a type t who is indifferent between $s^*(t) = t$ and $s^*(t) = \bar{s}$, i.e., for whom $1 + P(\bar{s}) = P(t)$. Note also that the sources of the pressure $P(s)$ can be divided into two sections – those that compose the rightmost uniform part, and those that lie to the left of this uniform part. The sources of the first section impose the same pressure on the type at the left edge of the rightmost uniform part and on the type at the right edge of this uniform part (due to symmetry). The sources of the second section impose more pressure on the latter, because this type is farther away from the norm. Together with the fact that D is the same for both types, this implies that $1 + P(\bar{s}) < P(t)$ for this latter type, in contradiction to the assumption that this type chooses $s^*(t) = t$. Since a rightmost and detached uniform part cannot exist this implies that no detached uniform part can exist to the right of $\bar{s}$. A similar argument applies to the left of $\bar{s}$ and hence we have shown that the only uniform part that can exist is attached to $\bar{s}$.

Finally, we need to show that this uniform part can be written as $[\bar{s}-y, \bar{s}+y] \cap [-1,1]$ for some y, which boils down to showing that it cannot be asymmetric if it does not touch any of the edges of the type distribution. I.e., this part cannot be $[\bar{s}-y_1, \bar{s}+y_2] \subset [-1,1]$ where $y_1, y_2 > 0$ and $y_1 \neq y_2$. Suppose to the contrary that this case holds. Then the aggregate pressure $P(s)$ is given by:

$$P(\sigma) = \begin{cases} Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(\sigma+y_1)^{\beta+1}+(y_2-\sigma)^{\beta+1}}{\beta+1} & \text{if } -y_1 \leq \sigma \leq y_2 \\ Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(y_2-\sigma)^{\beta+1}-(-y_1-\sigma)^{\beta+1}}{\beta+1} & \text{if } \sigma < -y_1 \\ Kx|\sigma|^{\beta} + K\frac{1}{2}\frac{(\sigma+y_1)^{\beta+1}-(\sigma-y_2)^{\beta+1}}{\beta+1} & \text{if } \sigma > y_2 \end{cases} \tag{35}$$

where $x = \frac{y_1+y_2}{2}$. *Moreover, both the type* $t_1 = \bar{s} - y_1$ *and the type* $t_2 = \bar{s} - y_2$ *are indifferent between* $s^*(t) = t$ *and* $s^*(t) = \bar{s}$. *Hence it must hold simultaneously that* $1 + P(0) = P(-y_1)$ *and* $1 + P(0) = P(y_2)$, *i.e.,* $P(-y_1) = P(y_2)$. *Substituting* $\sigma = -y_1$ *and* $\sigma = y_2$ *in equation (35) we get*

$$Kxy_1^\beta + K\frac{1}{2}\frac{(y_2+y_1)^{\beta+1}}{\beta+1} = Kxy_2^\beta + K\frac{1}{2}\frac{(y_2+y_1)^{\beta+1}}{\beta+1}$$
$$\Rightarrow y_1^\beta = y_2^\beta$$

which contradicts $y_1 \neq y_2$. ■

Lemma 18 *Suppose* $\beta \leq 1$. $K_{\min}(|\bar{s}|)$ *is weakly decreasing in* $|\bar{s}|$.

Proof. *We start with the case* $\beta < 1$. *First note that* $K_{\min}$ *is never found on the border between the regions (1) and (2),*[61] *since* $\frac{d(1/K)}{dy}|_{y\to^+ s_l}$ *is strictly greater (unless* $s_l = 0$*) than* $\frac{d(1/K)}{dy}|_{y\to^- s_l}$. *We can therefore rewrite equation (32) as a function of* $\bar{s}$ *for the two regions and differentiate* $1/K$ *w.r.t.* $\bar{s}$. *This yields*

$$\frac{d(1/K)}{d\bar{s}} = \begin{cases} 0 \text{ if } y \leq \bar{s}+1 \\ -\frac{y^\beta}{2} + \frac{1}{2}(\bar{s}+y+1)^\beta - \frac{1}{2}(\bar{s}+1)^\beta \text{ if } y \in [\bar{s}+1, \min\{y_{\max}(\bar{s}), 1-\bar{s}\}] \end{cases} \quad (36)$$

$$\frac{d^2(1/K)}{d\bar{s}^2} = \begin{cases} 0 \text{ if } y \leq \bar{s}+1 \\ \frac{1}{2}\beta(\bar{s}+y+1)^{\beta-1} - \frac{1}{2}\beta(\bar{s}+1)^{\beta-1} \text{ if } y \in [\bar{s}+1, \min\{y_{\max}(\bar{s}), 1-\bar{s}\}] \end{cases} \quad (37)$$

Note that $\frac{d(1/K)}{d\bar{s}}|_{y\to^+\bar{s}+1} = \left(2^{\beta-1}-1\right)(\bar{s}+1)^\beta < 0$ *and* $\frac{d^2(1/K)}{d\bar{s}^2} \leq 0$. *These results imply that* $\frac{1}{K}(y)$ *is constant in* $\bar{s}$ *in the first region and is strictly decreasing in* $\bar{s}$ *in region (2) (note that this does not violate the continuity of* $\frac{1}{K}(y)$ *as can be verified by plugging* $y = s_l$ *in equation (32)). Hence, since we have been analyzing the case of* $\bar{s} \leq 0$, *more generally* $K(y)$ *is weakly decreasing in* $|\bar{s}|$. *In particular* $K_{\min}$ *is weakly decreasing in* $|\bar{s}|$ *– it stays constant if* $K_{\min}$ *is achieved in region (1) both before and after the change in* $|\bar{s}|$, *and is strictly decreasing if* $K_{\min}$ *is achieved in region (2) after the change in* $|\bar{s}|$.

Now for the case $\beta = 1$. *Plugging* $\beta = 1$ *into equation (32) we get that both regions are independent of* $\bar{s}$. *Hence,* $K_{\min}$ *is independent of* $\bar{s}$. ■

Proof of Proposition 4

Lemma 15 implies that for any $\bar{s} \in [-1,1]$, one can construct $S(y)$ such that all types $t \in [\bar{s}-y, \bar{s}+y] \cap [-1,1]$ for some $y \leq y_{\max}(\bar{s})$ speak their minds while the rest choose $\bar{s}$, and $S(y)$ with $\bar{s}$ form a single norm equilibrium, if a suitable value of K is chosen. Moreover, this lemma says that $K(y)$, the value for which this single norm equilibrium exists for a given y, is either U-shaped or W-shaped as a function of y when $\beta < 1$; and $K(y)$ is strictly decreasing in y with a min point at $y = y_{\max}$

[61] These regions are defined in the proof of Lemma 15.

when $\beta = 1$. When $y \to 0$ we have

$$\lim_{y\to 0} 1/K = \lim_{y\to 0} \left\{ (1-y)\, y^{\beta} + \left(2^{\beta} - 1\right) \frac{y^{\beta+1}}{\beta+1} \right\} = 0,$$

so that $K(y) \to \infty$. Let $K_{\min}(|\bar{s}|)$ denote the minimal value of $K(y)$. It thus immediately follows that for $K \geq K_{\min}(|\bar{s}|)$ there exists a fix point y while for $K < K_{\min}(|\bar{s}|)$ there does not. This proves the if part of statement (1). As for the only if part of the statement, note that Lemma 17 implies that in any single norm equilibrium, all types $t \in [\bar{s} - y, \bar{s} + y] \cap [-1, 1]$ for some $y < 1 + |\bar{s}|$ speak their minds while the rest choose $\bar{s}$. It thus suffices to show that if such an equilibrium exists for some $y > y_{\max}(\bar{s})$, then still $K \geq K_{\min}(|\bar{s}|)$. For $\beta < 1$ this is proved in Lemma 16. For $\beta = 1$ we know from Lemma 12 that $y_{\max} = 1$. Then, when $y > y_{\max} = 1$, no K can sustain a single norm equilibrium at $\bar{s}$. This can be seen by setting $\beta = 1$ and letting $s \to^{+} \bar{s}$ in equation 25, and noting that, for $y > 1$, $\bar{s}$ is not the global min point of P and so cannot be the norm given that D is a step function. As for statement (2) of the proposition, the fact that $K_{\min}$ is weakly decreasing in $|\bar{s}|$ follows directly from Lemma 18.■

C.3.4 Proof of Proposition 5

The proof of the proposition builds on a few auxiliary lemmas, and on expressions within these lemmas, that are outlined first. The actual proof of the proposition follows after the lemmas.

Lemma 19 *Suppose $\beta \leq 1$. Suppose in some generation i there exists a cutoff distance from the norm y_i, such that all types in that generation that fulfill $|t - \bar{s}| > y_i$ declare the norm and all types fulfilling $|t - \bar{s}| \leq y_i$ speak their minds and that $y_i \leq y_{\max}(\bar{s})$. Then there exists a cutoff y_{i+1} in the next generation, such that all types that fulfill $|t - \bar{s}| > y_{i+1}$ declare the norm and all types that fulfill $|t - \bar{s}| \leq y_{i+1}$ speak their minds. Furthermore y_{i+1} is an increasing function of y_i.*

Proof. *When $y_i \leq y_{\max}(\bar{s})$ then by Lemma 4 P is increasing with distance from $\bar{s}$. Since D is a fixed cost it implies that types sufficiently far from $\bar{s}$ declare $\bar{s}$ and types sufficiently close declare their type (note that this cutoff may be such that all types declare their type). By Lemma 14 we know that if the cutoff type $t = \bar{s} + y_{i+1}$ is such that $\bar{s} - y_{i+1} < -1$ then type $t = -1$ strictly prefers stating her type. This implies that we only need to focus on the indifferent type $t > \bar{s}$. The indifferent type (which we define as $t_c \equiv \bar{s} + y_{i+1}$) is such that*

$$L(t_c, t_c) = P_{i+1}(t_c) = P_{i+1}(\bar{s}) + D(t_c, \bar{s}) = L(t_c, \bar{s}).$$

Define

$$F \equiv D(t_c, \bar{s})/K + P_{i+1}(\bar{s})/K - P_{i+1}(t_c)/K = 0.$$

Then $F = 0$ implicitly gives us y_{i+1} as a function of y_i). For a given y_i, F can take

one of the following forms:

$$F = \tag{38}$$

$$\begin{cases} F_1 \equiv \frac{1}{K}+\frac{1}{2}\frac{(\bar{s}+1)^{\beta+1}+(y_i)^{\beta+1}}{\beta+1}-\frac{1}{2}\left[(1-y_i-\bar{s})\,(y_{i+1})^{\beta}+\frac{(\bar{s}+y_{i+1}+1)^{\beta+1}+(y_i-y_{i+1})^{\beta+1}}{\beta+1}\right] \text{ if } y_i \geq y_{i+1}, \bar{s}-y_i < -1 \\ F_2 \equiv \frac{1}{K}+\frac{y_i^{\beta+1}}{\beta+1}-\left[(1-y_i)\,(y_{i+1})^{\beta}+\frac{1}{2}\frac{(y_{i+1}+y_i)^{\beta+1}+(y_i-y_{i+1})^{\beta+1}}{\beta+1}\right] \text{ if } y_i \geq y_{i+1}, \bar{s}-y_i \geq -1 \\ F_3 \equiv \frac{1}{K}+\frac{1}{2}\frac{(\bar{s}+1)^{\beta+1}+(y_i)^{\beta+1}}{\beta+1}-\frac{1}{2}\left[(1-y_i-\bar{s})\,(y_{i+1})^{\beta}+\frac{(\bar{s}+y_{i+1}+1)^{\beta+1}-(y_{i+1}-y_i)^{\beta+1}}{\beta+1}\right] \text{ if } y_i \leq y_{i+1}, \bar{s}-y_i < -1 \\ F_4 \equiv \frac{1}{K}+\frac{y_i^{\beta+1}}{\beta+1}-\left[(1-y_i)\,y_{i+1}^{\beta}+\frac{1}{2}\frac{(y_{i+1}+y_i)^{\beta+1}-(y_{i+1}-y_i)^{\beta+1}}{\beta+1}\right] \text{ if } y_i \leq y_{i+1}, \bar{s}-y_i \geq -1 \end{cases}$$

Note that when $\bar{s}-y_t \to -1$ then $F_1 = F_2$ and $F_3 = F_4$; that when $y_{i+1} \to y_i$ then $F_1 = F_3$ and $F_2 = F_4$; and finally that when $\bar{s}-y_i \to -1$ and $y_{i+1} \to y_i$ then $F_1 = F_3 = F_2 = F_4$. Hence, since each of F_1, F_2, F_3 and F_4 is continuous then F is a continuous function and hence y_{i+1} is a continuous function of y_i. This implies that, if y_{i+1} is an increasing function y_i for each of F_1, F_2, F_3 and F_4, then y_{i+1} is an increasing function of y_i also globally. By the implicit function theorem we have

$$\frac{dy_{i+1}}{dy_i} = -\frac{F_{y_i}}{F_{y_{i+1}}}.$$

Note that the bracket in each F equals $P(s)\,|_{s=y_{i+1}}$, which implies that

$$F_{y_{i+1}} = -\frac{dP}{dy_{i+1}} = -\frac{dP}{ds}|_{s=y_{i+1}}, \tag{39}$$

which we know is negative by Lemma 4. Hence, if F_{y_i} is positive then $\frac{dy_{i+1}}{dy_i}$ is positive.

$$F_{y_i} = \begin{cases} \frac{1}{2}(y_i)^{\beta}+\frac{1}{2}(y_{i+1})^{\beta}-\frac{1}{2}(y_i-y_{i+1})^{\beta} & \text{if } y_i \geq y_{i+1}, \bar{s}-y_i < -1 \\ y_i^{\beta}+y_{i+1}^{\beta}-\frac{1}{2}(y_{i+1}+y_i)^{\beta}-\frac{1}{2}(y_i-y_{i+1})^{\beta} & \text{if } y_i \geq y_{i+1}, \bar{s}-y_i \geq -1 \\ \frac{1}{2}y_i^{\beta}+\frac{1}{2}y_{i+1}^{\beta}-\frac{1}{2}(y_{i+1}-y_i)^{\beta} & \text{if } y_i < y_{i+1}, \bar{s}-y_i < -1 \\ y_i^{\beta}+y_{i+1}^{\beta}-\frac{1}{2}(y_{i+1}+y_i)^{\beta}-\frac{1}{2}(y_{i+1}-y_i)^{\beta} & \text{if } y_i > y_{i+1}, \bar{s}-y_i > -1 \end{cases}$$

From this expression one can see that F_{y_i} is strictly positive on all rows: the first and third row trivially follow from $\frac{1}{2}(y_i)^{\beta} > \frac{1}{2}(y_i-y_{i+1})^{\beta}$ and the second and fourth row follow since $\frac{1}{2}y_i^{\beta}+\frac{1}{2}y_{i+1}^{\beta} \geq \frac{1}{2}(y_{i+1}+y_i)^{\beta}$ and $\frac{1}{2}y_i^{\beta} > \frac{1}{2}(y_i-y_{i+1})^{\beta}$. ■

Lemma 20 *Suppose $\beta \leq 1$. Then:*

1. *$y_{\max}(\bar{s})$ (from Lemma 4) is weakly increasing in $|\bar{s}|$.*
2. *Let $K(y)$ be implicitly given by equation (31) and let $\tilde{y}$ denote an implicit solution to this equation for a given value of K. Then if $K'(\tilde{y}) > 0$, $\tilde{y}$ is weakly increasing in $|\bar{s}|$, and if $K'(\tilde{y}) < 0$, $\tilde{y}$ is weakly decreasing in $|\bar{s}|$*

Proof. *$y_{\max}(\bar{s})$ is the maximum value of y such that $P(s)$ is monotonically increasing in $|s-\bar{s}|$. In Lemma 4 we show that it is unique for a given $\bar{s}$, such that $P(s)$ is monotonically increasing if and only if $y \leq y_{\max}(\bar{s})$. For $\beta = 1$ we know from Lemma 12 that $y_{\max} = 1 \ \ \forall \bar{s}$. For $\beta < 1$ we will show that $y_{\max}(s_l)$ is decreasing in s_l (recall that $s_l \equiv \bar{s}+1$), which is equivalent to the first statement in the lemma. Suppose that s_l is given, and that $y = y_{\max}(s_l)$. It follows then that $\exists s \in [-1,1]$ such that $P'(s) = 0$. If we then increase s_l by some ϵ while keeping $y = y_{\max}(s_l)$, we get by equation (28) that $\exists s \in [-1,1]$ such that $P'(s) < 0$, implying that $P(s)$ is not monotonically increasing in $|s-\bar{s}|$ for any $y \leq y_{\max}(s_l)$. This means that $y_{\max}(s_l+\epsilon) < y_{\max}(s_l)$, i.e., $y_{\max}(\bar{s})$ is increasing in $|\bar{s}|$ as in statement (1).*

2) Equation (31) depicts the function $K(y)$ in region (2) (as defined in Lemma 15). From the proof of Lemma 15 we know that if $\beta < 1$ then $K(y)$ is weakly decreasing in $|\bar{s}|$ and if $\beta = 1$ then $K(y)$ is constant in $|\bar{s}|$, and this holds in particular for region (2). It thus follows that, for a given value of K, any implicit solution $\tilde{y}$ for which $K'(\tilde{y}) > 0$ is weakly increasing in $|\bar{s}|$, and any implicit solution $\tilde{y}$ for which $K'(\tilde{y}) > 0$ is weakly decreasing in $|\bar{s}|$. ■

Proof of Proposition 5

1) Recalling that $F = 0$ in equation (38) implicitly gives us $y_{i+1}(y_i)$, we can see in that equation that when $y_i = 0$, the only way for F to equal zero is to have $F = F_4 = 1/K - y_{i+1}^{\beta}$, implying that $y_{i+1}(0) > 0$.[62] Lemma 19 further shows that y_{i+1} is an increasing function of y_i. If $K < K_{\min}(|\bar{s}|)$, we know from Lemma 15 that no steady state exists. Otherwise, if $K \geq K_{\min}(|\bar{s}|)$, then by Lemma 15 we know that a steady state exists (at least one). Next, note that F in equation (38) is strictly decreasing in K (this applies to F_1, F_2, F_3 and F_4). This implies that $F_K < 0$, which together with $F_{y_{i+1}} < 0$ (see equation 39) implies that $\frac{dy_{i+1}}{dK} = -\frac{F_K}{F_{y_{i+1}}} < 0$, i.e., that the function $y_{i+1}(y_i)$ goes down when K increases. This means that when $K < K_{\min}(|\bar{s}|)$, the function $y_{i+1}(y_i)$ always stays above the 45 degree line (i.e. the line that implies $y_{i+1} = y_i$); when $K = K_{\min}(|\bar{s}|)$ the function $y_{i+1}(y_i)$ is tangent to the 45 degree line, and when $K > K_{\min}(|\bar{s}|)$ the function $y_{i+1}(y_i)$ crosses the 45 degree line at least once. It thus follows that when $K = K_{\min}(|\bar{s}|)$, any steady state would not be stable, as there can be no convergence to it from the right. Furthermore, if $K > K_{\min}(|\bar{s}|)$, it implies together with $y_{i+1}(0) > 0$ that there must be at least one stable steady state, as there is at least one point where the function $y_{i+1}(y_i)$ crosses the 45 degree line, starting above it and continuing below it. Denoting the leftmost stable steady state by y_{ss} and $\min\{y(K_{\min}(\bar{s}))\}$ by $y_{\min}(\bar{s})$ (note that $y(K_{\min}(\bar{s}))$ is unique if $K(y)$ is U-shaped and may have at most two solutions when it is W-shaped). Then we know that $y_{ss} \leq y_{\min}(\bar{s})$ because our analysis up till now implies that $y_{i+1}(y_{\min}(\bar{s})) < y_{\min}(\bar{s})$.[63] From $y_{i+1}(0) > 0$ we know that $y_{ss} \neq 0$, and since

[62] To see this note that when $y_i = 0$, F_4 and F_2 are the only relevant cases and that if $F = F_2$ then by construction it must be that $y_{i+1} = 0$ implying $F = F_2 \equiv 1/K \neq 0$, which contradicts $F = 0$.

[63] Note that $y_{\min}(\bar{s})$ is a steady state when $K = K_{\min}(|\bar{s}|)$, in which case $y_{i+1}(y_{\min}(\bar{s})) = y_{\min}(\bar{s})$. As K is further increased, $y_{i+1}(y_{\min}(\bar{s}))$ goes down.

$y_{ss} \leq y_{\min}(\bar{s})$, it follows that $x_{ss} \in]0,1[$.

2) Let now $K > K_{\min}(|\bar{s}|)$ and take a steady state, be it stable or unstable. To verify stability we need to compute dy_{i+1}/dy_i at the steady state – it is stable from both sides if and only if the derivatives are smaller than 1. To simplify calculations, note first that in steady states, where $y_{i+1} = y_i$, we get that $\frac{dF_1}{dy_i} = \frac{dF_3}{dy_i}$ and $\frac{dF_2}{dy_i} = \frac{dF_4}{dy_i}$, which means that we can work solely with F_3 and F_4, depending on the region of y, as defined in Lemma 15.[64] If the steady state falls in the first region, where $y < s_l$, then F_4 applies. There we have

$$\begin{aligned}\frac{dy_{i+1}}{dy_i} &= -\frac{F_{y_i}}{F_{y_{i+1}}} \\ &= -\frac{y_t^\beta + y_{t+1}^\beta - \frac{1}{2}(y_{t+1}+y_t)^\beta - \frac{1}{2}(y_{t+1}-y_t)^\beta}{-\left[(1-y_t)\beta y_{t+1}^{\beta-1} + \frac{1}{2}(y_{t+1}+y_t)^\beta - \frac{1}{2}(y_{t+1}-y_t)^\beta\right]} \\ &= \frac{2y_i^\beta - 2^{\beta-1}y_i^\beta}{\left[(1-y_i)\beta y_i^{\beta-1} + 2^{\beta-1}y_i^\beta\right]}\end{aligned} \tag{40}$$

which is strictly smaller than 1 iff

$$\begin{aligned}2y_i^\beta - 2^{\beta-1}y_i^\beta &< (1-y_i)\beta y_i^{\beta-1} + 2^{\beta-1}y_i^\beta \\ y_i &< \frac{\beta}{(2-2^\beta+\beta)}.\end{aligned}$$

One can verify that $\frac{\beta}{(2-2^\beta+\beta)}$ is the FOC solution in region (1) (to see this, one can equate the first part of equation (33) to 0 and solve for y). From Lemma 15 we know that this is the only local extremum in region (1) and that this is a minimum point. Hence, in this region, a steady state y_i is stable if and only if $\frac{dK}{dy}|_{y_i} < 0$. If instead the steady state falls in the second region, where $y > s_l$, then F_3 applies. There

$$\begin{aligned}\frac{dy_{i+1}}{dy_i} &= -\frac{F_{y_i}}{F_{y_{i+1}}} \\ &= -\frac{\frac{1}{2}y_t^\beta + \frac{1}{2}y_{t+1}^\beta - \frac{1}{2}(y_{t+1}-y_t)^\beta}{-\left[\left(1-\frac{y_t}{2}-\frac{(\bar{s}+1)}{2}\right)\beta y_{t+1}^{\beta-1} + \frac{1}{2}(\bar{s}+y_{t+1}+1)^\beta - \frac{1}{2}(y_{t+1}-y_t)^\beta\right]} \\ &= \frac{y_i^\beta}{\left[\left(1-\frac{y_i}{2}-\frac{(\bar{s}+1)}{2}\right)\beta y_i^{\beta-1} + \frac{1}{2}(\bar{s}+y_i+1)^\beta\right]}\end{aligned} \tag{41}$$

[64] Unless the steady state falls exactly at the border between the two regions, where $y = s_l$, in which case there is convergence to this steady state only from one side.

which is smaller than 1 iff

$$(1 - y_i - \bar{s})\,\beta y_i^{\beta-1} + (\bar{s} + y_i + 1)^{\beta} - 2y_i^{\beta} > 0.$$

This inequality (short of a factor of 1/2) corresponds to $d(1/K)/dy$ being positive in the second region, as can be seen in the second region of equation (33). That is, in this region too, a steady state y_i is stable if and only if $\frac{dK}{dy}|_{y_i} < 0$. Finally, we know that in steady states, equation (32) holds. If the steady state is in region (1) of this equation, then it is independent of $\bar{s}$. Otherwise the steady state is in region (2). Then part (2) of Lemma 20 says that if in a steady state y_i we have $K'(y_i) > 0$, then y_i is increasing in $|\bar{s}|$, and if we have $K'(y_i) < 0$, y_i is decreasing in $|\bar{s}|$. Therefore, in all stable steady states we get that y_i is weakly decreasing in $|\bar{s}|$, implying that the share of norm conformers $x_{ss}(|\bar{s}|)$ is weakly increasing in $|\bar{s}|$.

3) Since $K > K_{\min}(|\bar{s}|)$ is given, we know from the proof of statement (1) that there exists a stable steady state with a single norm $\bar{s}$ such that there is convergence to it from any $y_i < y_{ss}$. To show convergence to a stable steady state from the right, let $y_{conv} \equiv \min\{y_{uss}, y_{\max}(|\bar{s}|)\}$, where y_{uss} is the rightmost steady state in $[0, y_{\max}(|\bar{s}|)]$ that is unstable from both sides, if such a one exists. Suppose y_{uss} does not exist, so that $y_{conv} = y_{\max}(|\bar{s}|)$. Then either there is a unique, and stable, steady state y_{ss}, and therefore $y_{i+1} < y_i \;\; \forall y_i \in]y_{ss}, y_{\max}(|\bar{s}|)]$, implying convergence to y_{ss}; or, there may be steady states in $]y_{ss}, y_{\max}(|\bar{s}|)]$ that are unstable only from one side, in which case $y_{i+1} < y_i$ in their neighborhood, implying once again convergence to y_{ss}. Otherwise $y_{conv} = y_{uss}$, and the complete instability of y_{uss} implies that when $y_i \xrightarrow{-} y_{uss}$, $y_{i+1} < y_i$, and so there is convergence to a stable ss from any $y_i < y_{uss}$.[65]

4) Revisiting Lemma 20, part (1) of that lemma implies that $y_{conv}(|\bar{s}|)$ is increasing in $|\bar{s}|$ whenever $y_{conv} = y_{\max}(|\bar{s}|)$. If instead $y_{conv} = y_{uss}$, then it was shown in the proof to statement (2) of this proposition that $y_{conv}(|\bar{s}|)$ is weakly increasing in $|\bar{s}|$. This concludes the proof.■

C.4 Descriptive and prescriptive norms

C.4.1 Proof of Proposition 6

1) In the single norm equilibria in Proposition 2, P has the properties given by equation (11), whereby the norm is trivially the minimum point of social pressure. In the single norm equilibria in Proposition 4, $y \leq y_{\max}$ (see the proof of that proposition). By Lemma 4 we know that P is increasing in the distance from $\bar{s}$ whenever $y \leq y_{\max}$.

2) Follows from Lemma 5.■

[65] There may be two stable steady states to the left of y_{uss}, with convergence from small values of y_i to the first steady state and from large values of y_i to the second steady state, but this statement, and hence statement (3) of the proposition, holds in this case too.