Michaeli, Moti; Spiro, Daniel

Working Paper

Skewed norms under peer pressure: Formation and collapse

Memorandum, No. 15/2014

Provided in Cooperation with:
Department of Economics, University of Oslo

Suggested Citation: Michaeli, Moti; Spiro, Daniel (2014) : Skewed norms under peer pressure: Formation and collapse, Memorandum, No. 15/2014, University of Oslo, Department of Economics, Oslo

This Version is available at:
https://hdl.handle.net/10419/102053

MEMORANDUM

No 15/2014

Skewed Norms under Peer Pressure: Formation and Collapse

Moti Michaeli and Daniel Spiro

ISSN: 0809-8786

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.sv.uio.no/econ
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no
e-mail: frisch@frisch.uio.no

Last 10 Memoranda

No 14/14	Daniel Spiro *Resource Prices and Planning Horizons*
No 13/14	Johan Gars and Daniel Spiro *Uninsurance through Trade*
No 12/14	Moti Michaeli and Daniel Spiro *Powerty in China as Seen from Outer Space*
No 11/14	Ingvild Almås, Åshild Auglænd Johnsen and Andreas Kotsadam *Powerty in China as Seen from Outer Space*
No 10/14	Nico Keilman and Coen van Duin *Stochastic Household Forecast by Coherent Random Shares Prediction*
No 09/14	Mads Greaker, Michael Hoel and Knut Einar Rosendahl *Does a Renewable Fuel Standard for Biofuels Reduce Climate Costs?*
No 08/14	Karine Nyborg *Project Evaluation with Democratic Decision-making: What Does Cost-benefit Analysis Really Measure?*
No 07/14	Florian Diekert, Kristen Lund and Tore Schweder *From Open-Access to Individual Quotas:* *Disentangling the Effects of Policy Reform and* *Environmental Changes in the Norwegian Coastal Fishery*
No 06/14	Edwin Leuven, Erik Plug and Marte Rønning *Education and Cancer Risk*
No 05/14	Edwin Leuven, Erik Plug and Marte Rønning *The Relative Contribution of Genetic and Environmental Factors to Cancer Risk and Cancer Mortality in Norway*

Previous issues of the memo-series are available in a PDF® format at:
http://www.sv.uio.no/econ/english/research/memorandum/

Skewed Norms under Peer Pressure: Formation and Collapse*

Moti Michaeli† & Daniel Spiro‡

Memo 15/2014-v1
(This version June 2014)

Abstract

This paper shows that peer pressure may lead to dynamic convergence to a norm that is skewed with respect to preferences in society, yet is endogenously upheld by the population. Moreover, a skewed norm will often be more sustainable than a representative norm. This may explain the skewness of various social and religious norms. By furthermore interpreting a norm as a political regime, we show that biased regimes can be sustained even without the existence of a powerful group with coherent interests. We analyze the pattern by which political regimes collapse and relate it to contemporary revolutions and mass protests.

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

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

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; to follow religious customs to the same extent as their co-believers; and in order to avoid arguments most prefer not to state controversial

*We wish to thank Martin Dufwenberg, Tore Ellingsen, Juan Esteban, Bård Harstad, Charles Manski, Arie Michaeli, Kalle Moene, Manuel Oeschlin, Andrew Oswald, Gerard Roland, Moshe Shayo, 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, and Center for the Study of Rationality, the Hebrew University, motimich@gmail.com.

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

political opinions. Minimizing social pressure in these situations is often more complex than just following a clear social norm. For example, if a Muslim girl has one friend wearing the Burqa, another one wearing the Hijab, and a third one 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 behavior of others.

Meanwhile, when behaving in certain ways in public, people also indirectly affect others. For instance, college students who drink in fear of being unpopular, also indirectly strengthen the norm of drinking (see Centola et al., 2005). In modeling terms this means that – when each person declares a stance balancing her private opinion and the peer pressure – we get many stated opinions that on aggregate shape the social custom and hence what is considered to be normative. Such modeling lends itself to analyzing the emergence of an endogenous norm – a single mode of behavior actually followed by many in society. This norm is *descriptive*, reflecting a unique point along a continuum of possible stances, chosen by a significant number of individuals despite their heterogenous preferences.[1] This way, the very existence of a social norm is not simply assumed but is an equilibrium outcome contingent on people behaving according to it.

We address the issue of endogenous norm formation under peer pressure by analyzing the interaction between (infinitely) many individuals with heterogenous private opinions who can choose heterogenous public stances. To move ahead with this rather complex problem, we focus on the case where the disutility of an individual stating a different opinion than her private one increases *concavely* with the distance between these opinions. This represents the notion that once a person deviates a little from her own bliss point, further deviation hardly matters. Recent experimental research provides support for this assumption, showing 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).[2]

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

[1] 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). Our analysis contains both of these possibilities, where prescriptive norms are interpreted as actions that minimize social pressure. We focus on the former in the body of the paper and analyze the relation between these two kinds of norms in the appendix.

[2] One can of course also think of situations where a convex disutility is reasonable. But for issues of ideology and religion, concave disutility seems plausible, as suggested by Kendall et al. (2013). Moreover, since misrepresenting one's private opinion in public is related also to honesty and not just to ideology, the supportive evidence on cheating is relevant here too (Gneezy et al., 2013; Gino et al. 2010).

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

Our analysis presents the conditions under which a norm, and in particular a skewed one, is sustainable. Whenever a unique norm emerges, one of two distinctive types of societies will evolve. 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 – thus being alienated. That is, individual non-conformity arises if there is large misalignment between the individual's private opinion and the norm. As a result, a skewed norm will be less sustainable than a central norm, since it generates misalignment with many individuals. The same logic applies to the dynamic stability of norms in this type of society. Here norms will be undermined when private sentiments shift away from the norm.

In the second type of society, there are always people voicing their disagreement with the norm. Somewhat surprisingly, those will be the people with only minor disagreement with the norm. 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. They nevertheless do so because, lacking the ability to replace the norm, they would rather join it.[3] We call this type of society an *inverting society*, as public conformity and private conformity are inverted. In this society, the norm draws its strength from those who privately disagree with it. Hence, a norm that is skewed with respect to private sentiments can survive under *weaker* conditions than a non-skewed norm

[3] We implicitly assume that people do not have the option of refraining from declaring a stance. This is fairly standard in the literature (e.g., Kuran, 1989a; Granovetter, 1978; Bernheim, 1994; Manski and Mayshar, 2003; Kuran and Sandholm, 2008; Rubin, 2014). One can think of situations in which staying silent is either literally impossible (as in the case of choosing a headwear for a Muslim girl), or has the same peer effect as fully conforming (as in the case of passive obedience). Alternatively, one way of not declaring a stance would be by emigrating. In this case the implicit assumption is that emigration is too costly.

and will also be dynamically more stable. Furthermore, here we should expect a norm to be undermined by an increase in the share of people posing mild critique, following a shift in private sentiments to be *more aligned* with the norm.

This analysis also has implications for the sustaining and collapse of political regimes. One standard way of modeling regime collapse is to assume that individuals, while having heterogeneous private preferences, face the binary choice of either supporting the regime or protesting against it (Kuran, 1989a; Granovetter, 1978). If in equilibrium no one supports the regime, the interpretation is that the regime collapses. Our model provides a microstructure to such a setup. In our model, individuals, being heterogeneous in their private preferences, can choose the extent of support of the regime from a continuum. The strength of the regime (a.k.a. the norm) then depends on expressed public support, and the regime becomes weaker the more individuals criticize it and the harsher this critique is. Naturally, if a strong group with coherent interests exists, this can lead to additional clustering of supporters beyond that group. But we show that a regime can be strengthened and sustained even in the absence of such a group, i.e., even when private interests are fully heterogeneous. A biased political regime can then be seen as the counterpart of a skewed norm. As Kuran (1989a, 1989b, 1995) has pointed out, some regimes such as the former Soviet Union remain in power even though they do not represent the people's interest. 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. As Arendt (1964, p. 175) concludes, the "ideal of toughness, except, perhaps, for a few half-demented brutes, was nothing but a myth of self-deception, concealing a ruthless desire for conformity at any price".[4]

Our dynamic analysis further outlines what triggers a revolution, and describes the subsequent process. In an alienating society, revolutions will propagate from the outside towards the inside, following private sentiments becoming less aligned with the regime. As in Kuran (1989a), changes of private sentiments will first go unnoticed, but will eventually create a pocket of fierce opposition that will spark the revolution (which may often be initially violent). This will gradually induce more moderate objectors to speak their minds, until eventually no one follows the regime. This may be a reasonable description of the Iranian revolution in 1978-79, where there was a growing misalignment between the Shah and the religious sentiments in society (Razi, 1987). This revolution was initiated by the hardest opponents of the Shah, but then gained mass support by recruiting individuals with more moderate views (ibid). These dynamics may possibly also represent the evolution of the Russian revolution in 1917.

[4] As Arendt's (1964) analysis suggests, the Nazi regime had fairly extensive support within Germany itself. But once individuals were subject to opinions in other cultures they stopped supporting the Nazis. See also Cohen (2001) for a more thorough discussion and for opposing views.

Inverting societies, on the other hand, are more prone to sustain a biased political regime. However, 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 fairly agree with the regime, suggesting mild reform. This in turn will trigger new and gradually more fundamental suggestions, rejecting the regime from both sides of the political spectrum. Here revolutions will go from the inside out, resembling the sequence of events leading to the collapse of the Soviet Union and the communist regimes in eastern Europe (Pfaff, 2006) and possibly also to the recent collapse of Mubarak's regime in Egypt.[5]

In the next section we relate the paper to earlier research on social norms and on revolutions. In section 3 we present the individual decision problem under peer pressure, the concept of single norm equilibrium and a first result showing a class of societies that cannot maintain a social norm. Sections 4 and 5 analyze existence and dynamic stability of single norm equilibria in the two types of societies mentioned above. Section 6 discusses differences between the two types of society and, by interpreting the model as being about the formation of political regimes, discusses the revolutionary dynamics that may be observed in each. Section 7 concludes. In the appendix, we relate the equilibrium results to the existence of prescriptive norms and discuss how relaxing some model assumptions would change the main results of the model. All the formal proofs are relegated to a technical appendix at the end.

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 literature on social norms and conformity to peer pressure is confined to binary stances (e.g., Lindbeck et al., 2003; Brock & Durlauf, 2001; Lopez-Pintado & Watts, 2006; Kuran 1989a; Granovetter, 1978; Angeletos et al., 2007). Alternatively, when allowing continuous stances, it often takes the norm as exogenous (e.g., Bernheim, 1994). This naturally limits any investigation of endogenously formed skewed norms. Two exceptions are the models by Clark & Oswald (1998) and Michaeli & 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.[6] Since the existence of the norm is assumed in both these cases it is also harder to talk about a dynamic process

[5] Indeed in Egypt, the historically most extremist opponents of the regime (the Muslim brotherhood) were initially absent from the streets. Furthermore, like our model suggests, the "moderates" 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).

[6] Clark & Oswald do formulate a similar problem to ours, where the individual is affected by the actions of all others (p.144), but do not solve it.

that strengthens or undermines the norm. Furthermore, this limits the possibility to relate descriptive and prescriptive norms to each other, as we do in this paper.

Another way of sorting models of social pressure is along the two dimensions of (i) whether a person is punished for her private preference or for her behavior and (ii) whether the social pressure is formed by what individuals believe to be right or by their actual behavior. In this paper, pressure is applied to behavior and arises from behavior. Along the first dimension, in most of the previously mentioned papers on social pressure, individuals are punished for their behavior too. The alternative, when people are punished for their private preferences, has been pursued for instance by Bernheim (1994) and by Bénabou & Tirole (2006). This leads to a signalling game, where people are trying to be perceived as being of certain types, and try to reveal the true types of others. Along the second dimension, models of pressure that is formed by people's behavior typically concern situations of ideology, religion or more generally – situations where there is a true disagreement about what is right and what is wrong. The pressure can then be modeled either as pairwise peer pressure (Manski and Mayshar, 2003; Kuran & Sandholm, 2008) or as a cost for deviating from a descriptive norm that reflects the average behavior in society (Michaeli & Spiro, 2014). The other alternative is that pressure increases in the distance from what individuals believe to be right. When there is consensus about the right thing to do (e.g., being polite or working hard), it means that a prescriptive norm exists (McAdams, 1997; Cialdini et al., 1991). This is the case for example in models of status or work effort (Kandell & Lazear, 1992; Clark & Oswald, 1998; Dufwenberg and Lundholm, 2001).[7]

The papers that are most closely related to ours from a modeling perspective are Manski & Mayshar (2003) – where the choice of the number of children of one person depends on the choices of others – and Kuran & Sandholm (2008), who analyze the integration speed of groups with different preferences. In these papers, just like in ours, pressure arises between all pairs of individuals in society and is applied to the behavior of an individual, rather than her private preference. What these two papers have in common, while differing from us, is that they use 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. Furthermore, there can be no clustering of opinions (i.e., there can be no descriptive norms) unless the choice set is discrete. Likewise, 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

[7] 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 talk about a descriptive norm and a prescriptive norm in the same setup. We relate these two notions to each other, conceptually and analytically, in more detail in the appendix.

in a formal way. Although he argues convincingly that skewed equilibria can arise when the number of individuals is small, it is hard to know if this result applies more generally, i.e., when the number of individuals is larger and with other functional forms. Furthermore, there is no treatment of whether skewness is more or less stable than non-skewness.

Our paper also relates to the literature on revolutions and sustainability of political regimes. Following Tanter & Midlarsky (1967), there are two categories of revolutions. Firstly, *coup detats,* performed by elites or a competing party. Examples of these are plentiful in both Africa and Latin America. Typically, these are modeled by assuming the existence of an elite group (e.g. Acemoglu & Robinson, 2001). If several groups within a country exist, this also relates to civil wars (see Blattman & Miguel, 2010 for a survey), in which the groups' endogenously chosen fighting effort determines the probability of outcomes through a contest success function (e.g. Hirshleifer, 1988; Garfinkel, 1990; Skaperdas, 1992). The assumption that the group exists and makes collective decisions on behalf of its members is important in this setting, as it gets around the free rider problem in collective action.[8] This assumption may be plausible if the revolting group is small with sufficiently aligned preferences (see Goldstone, 1994, for a discussion).

The second category, which is more related to our paper, is labeled by Tanter & 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. These often lead to major changes of the political system, from one driven by elites to one having more popular support. 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 & Kitt, 1950; Festinger, 1954; Davies 1959; Davies 1962; and more recently for instance Tarrow, 1998; McAdam et al., 2001; Almer et al., 2013). But many popular protests also contain clear ideological or religious motives (Esteban & Ray 2011). For instance, the Iranian revolution (Razi, 1987), the revolutions in Eastern Europe (Lohmann, 1994), the rise of radical Islam (Beck, 2009) and social movements in Western Europe (Kriesi et al., 1992). Many important insights

[8] See Olson (1971), Tullock (1971) for an early treatment and, e.g., Oliver & Marwell, 1988; Esteban, 2001; Esteban & Ray, 2001), for more recent work on collective action. The chief decision maker in this class of models and in our model (and in the models of Kuran, 1989a; Granovetter, 1978; and Naylor, 1989) is the individual and not the group. The difference is, however, that the collective action literature focuses on individual homogeneity where individuals have direct incentives and disincentives for action while we focus on individual heterogeneity and where incentives for action come from some form of political motivation or personal economic gains. The collective action approach has also been used in various game theoretic settings to explain cultural diversity (Greif, 1994). Here individuals rationally expect certain behavior by others in equilibrium. This leads them to behave similarly. In our setting beliefs about others' behavior is rationally expected in equilibrium too, but the individuals are heterogenous in the private preferences.

regarding these environments can be gained by analyzing a simple model where individuals have the binary choice between giving support to the current regime (which possibly includes silent support) and protesting (see Kuran, 1989a, 1989b, 1995; Granovetter, 1978; Naylor, 1989; and more recently, for instance, Rubin, 2014).[9] But the binary approach has some limitations too, which are addressed by our paper. We discuss these additional insights at length and relate them to various real social movements and revolutions when presenting our results about revolutions in Section 6.

Our model also relates to research on endogenous preferences (e.g. Bisin & Verdier 2001; Kuran & Sandholm, 2008; Bowles, 1998; Roland, 2004). In the appendix, we discuss how alternative modeling of the dynamic process would affect our results.

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. I.e., some private preference, ideology or opinion, referred to also as the individual's *type*. For example, one can think of t as a position on a political scale. Each individual has then to publicly declare a chosen 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 increases in the distance between that stance and the individual's type, representing the cognitive dissonance or displeasure felt by her.

$$D = D\left(|t - s(t)|\right),\ D'(\cdot) \geq 0$$

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 is

$$p = p\left(|s - s'|\right),\ p'(\cdot) \geq 0.$$

Such pressure arises between each pair of individuals. This means that, given a set of stances in society $S \sim s'(\tau)$, the aggregated pressure (P) felt by an individual declaring some stance s is given by

$$P(s; S) \equiv E\left[p\left(|s - S|\right)\right] = \int_{t_l}^{t_h} p\left(|s - s'(\tau)|\right) f(\tau)\, d\tau, \tag{1}$$

[9] Note that in Rubin's (2013) paper the individual has a binary decision to support or not to support the regime, but the political regime itself can choose a more popular political policy (on a continuum) to avoid social unrest. However, the existence of a political regime is taken as exogenous.

where t_l and t_h are the borders of the distribution of types and $f(\tau)$ is the probability density function of types, which is assumed to be continuous. $s(\tau)$ is the stance taken by type τ.[10]

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

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

This formulation implies that the individual takes the distribution of stances S in society as given. Of course, in equilibrium the statement of the individual is in itself part of S, but we assume that there are sufficiently many individuals for each one *not* to take into account how her stance affects others' stances, and how that feeds back into affecting her.

Finding an equilibrium distribution of stances requires solving a fixed point problem, whose solution is a complete mapping from t to $s^*(t)$, where

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

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

That is, each individual choose 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. This is not an easy problem to solve under general conditions. Being interested in studying the emergence of a norm in society and in the conditions under which this norm may be skewed, 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 skewed.*

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 Section A we discuss the implication of the model for the existence of *prescriptive norms*, which put the focus on stances that are approved in society (i.e., reduce social pressure) yet are not necessarily followed in practice. Being interested in issues of ideology or religion, where there are truly differences of opinion with respect to what is the right thing to do, our view is that real descriptive norms is the most applicable modeling choice.

[10] 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 & Sandholm (2008).

In order to study the emergence of a single norm, we will confine our analysis to equilibria fulfilling the following condition.

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. To be up-front, the single norm equilibrium is not the only one that may exist in the this model, as more than one norm may arise. 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

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

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

(with $\alpha, \beta > 0$), and will restrict our attention to $\alpha < 1$, i.e., when cognitive dissonance is concave. As mentioned in the introduction, the assumption of $\alpha < 1$ has support in recent experimental research. Moreover, while D can be convex in some issues, for issues of ideology a concave disutility seems plausible and is also suggested by Kendall et al. (2013) following voters' attitudes. Finally, K represents the relative weight of the peer pressure, and so captures the extent to which individuals care about social pressure.

We start our analysis by showing in which societies single norm equilibria will not exist.

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

The proof of the lemma 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 $\beta > 1$, P will be convex in a neighborhood around the norm and will have a derivative of zero at the norm itself.[11] But with a zero derivative it becomes pointless to state the norm exactly, as a small deviation in the direction of one's private opinion reduces the dissonance without increasing the pressure.

The lemma implies that we can rule out societies where $\beta > 1$ from containing single norm equilibria. For $\beta \leq 1$ two distinctly different societies, which can contain a single norm equilibrium, emerge. The two alternatives are exhaustive (no other forms of single norm equilibria can exist) and they are treated in great detail in the next two sections. The first type of society is one that endogenously induces

[11] 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).

conformity by those who privately fairly agree with the norm, while alienating those who privately disagree with it strongly (covered in section 4). This happens if $\beta < \alpha$ and will be illustrated by assuming that p is a step function ($\beta \to 0$). The second type of society is one that endogenously induces conformity by those who privately dislike the norm the most (covered in section 5). This happens if $\alpha < \beta < 1$ and will be illustrated assuming that D is a step function ($\alpha \to 0$).[12] The usage of a step function in each case is made for brevity and to make analytical headway but does not drive the results.[13]

We will furthermore assume initially 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 skewed 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 later sections 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 very concave. 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 $\alpha \in]0, 1[$. A first useful result then follows.

Lemma 2 *Suppose that p is given by (8), D is given by (5) with $\alpha \in]0, 1[$ and that a single norm $\bar{s}$ exists and is stated by a share x of the population, while the rest speak their minds.*[14] *Define*

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

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

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

[12] To avoid technicalities, we will not analyze in this paper the special cases of $\alpha = \beta$ and of $\beta = 1$.

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

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

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. 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, when the only way to avoid being pressured by someone is to fully agree with her, 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 the norm relative to any other possible statement. 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 – the type closer to the norm feels lower dissonance. 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 3 *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).[15] This mapping also implies that, when holding y fixed, x is decreasing in $\bar{s}$.

A static equilibrium of the model is essentially a fixed point defined by a triplet $(x, y, \bar{s})$ that satisfy Lemma 2 and Lemma 3 simultaneously. The conditions for the existence of such an equilibrium are presented in the following proposition.

Proposition 1 *Suppose that individual pressure is according to (8) and D is given by (5) with $\alpha < 1$. 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 $K \geq K_{\min}(|\bar{s}|)$.*

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

This proposition expresses three main results, which hold also beyond the step function case. Firstly, that under very concave individual pressure there exist single norm equilibria whenever individuals care sufficiently about social pressure – K has to be greater than $K_{\min}(|\bar{s}|)$. Secondly, that in these equilibria the norm may be skewed. Thirdly, that the more skewed the norm is, the larger is the K needed to sustain it in equilibrium.[16] This last result is a key result. It essentially says that in order to uphold a skewed 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 skewed there are more private opinions further away from the norm and hence more potential deviators. To sustain the norm this has to be compensated

[15] If the norm is skewed, 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 $1 + |\bar{s}|$, the distance from the norm to the most extreme type in society, then everyone conforms to the norm, implying that $x = 1$.

[16] With $K_{\min}(0) = 1$ and $K_{\min}(1) = 2^{\alpha}$.

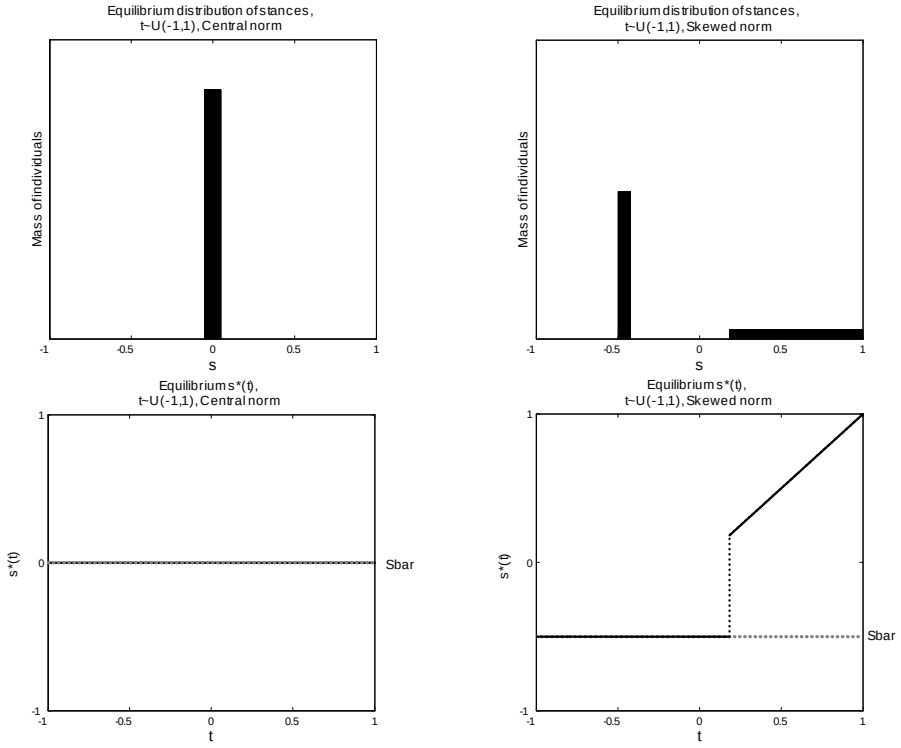

Figure 1: The left graphs show the distribution of stances and $s^*(t)$ in equilibrium with a central norm ($\bar{s} = 0$). The right graphs show the distribution of stances and $s^*(t)$ in equilibrium with a skewed norm ($\bar{s} = -0.5$). In all figures $\beta = 0.01$, $\alpha = 0.9$ and $K = 1.2$.

for by a heavier weight of pressure. This can also be seen in Lemma 3, which states that given y the share of norm followers falls with skewness 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 skewed 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.[17]

It is interesting 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 is repeated until a steady state is reached.[18] Let i indicate the period of the dynamic process (representing a period or a generation). Then an individual of 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. To get a sense of the equilibrium dynamics it is instructive to revisit Lemma 2. It is only a partial equilibrium result, but it implies that if a share of the population makes the same statement (a norm exists) in one period, then in the next period it will be optimal for those close to the norm to state the norm and for those far from it to speak their minds. Hence, given that a norm exists in one period, this will create alienation in the next period, and this alienation will be re-created in later periods too. The question then is whether this process converges to a stable single norm steady state or not. The answer is given in the following proposition.

Proposition 2 *Consider the dynamic model in (13) with p being a step function as in (8) and D as given in (5) with $\alpha < 1$. Then:*

[17] More precisely, Lemma 2 and Proposition 1 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.

[18] 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, by assuming that private preferences are not affected by the norm we essentially make it harder to sustain a norm than it would be otherwise. At any rate, in Section B we discuss the extent to which our results should hold under alternative dynamic assumptions too.

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* $x_{ss}(|\bar{s}|) > 0$ *of the population declare the norm.*

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

3. *Consider a norm* $\bar{s}$ *and suppose* $K > K_{\min}(|\bar{s}|)$. *Let* x_i *denote the share of norm followers in period* i. *Then there exists a value* $x_{conv}(|\bar{s}|)$ *such that if* $x_i > x_{conv}(|\bar{s}|)$, *there is convergence to a stable single norm steady state where a share* $x_{ss}(|\bar{s}|) > x_{conv}(|\bar{s}|)$ *state* $\bar{s}$. *Otherwise, if* $0 \leq x_i < x_{conv}(|\bar{s}|)$, *there is convergence to a stable steady state where each type speaks her mind* ($x_{ss}(|\bar{s}|) = 0$).[19]

4. $x_{conv}(|\bar{s}|)$ *increases in* $|\bar{s}|$ *and decreases in* K.

The proposition highlights that if $K > K_{\min}(|\bar{s}|)$ 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_{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 with, then the initial norm cannot be sustained at all, no matter how many declare it initially.[20]

With respect to the properties of dynamic convergence, Lemma 2 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 demonstrated in Figure 2 and is the main building block for proving Proposition 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

[19] For one specific value of K, there may exist an $\dot{x} < x_{conv}(|\bar{s}|)$, which is stable only with respect to convergence from above. The statement treats this special case as one where x_i, upon reaching $\dot{x}$, only passes through it and continues to $x_{ss} = 0$.

[20] 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 $x_{ss} < x_i \leq 1$).

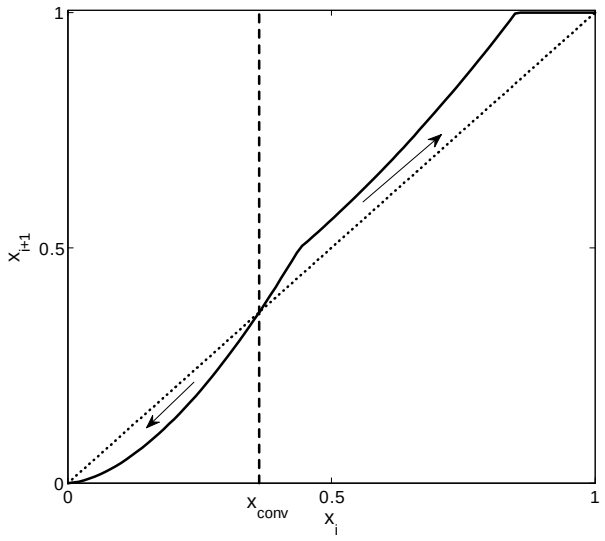

Figure 2: A phase diagram showing convergence to a single norm steady state with $\bar{s} = -0.5$ for p being a step function, $\alpha = 0.6$ and $K = 1.5$. The dotted line depicts the diagonal where $x_{i+1} = x_i$, the solid line depicts the intertemporal dynamics $x_{i+1} = f(x_i)$. The vertical line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

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.

Apart from convergence, the proposition also highlights the effect of the skewness of the norm. Parts (1) and (3) of the proposition imply that a skewed 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. 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 no longer represents 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 skewness 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 skewed norm *can* exist, the more skewed it is, the less magnetic it is, unless it is compensated for by a larger K. Hence, skewed 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 cohesiveness in society is falling with skewness.

5 Inverting societies

This section deals with the case where the dissonance function D is very concave. Following the terminology of Michaeli & Spiro (2014), we can label individuals as very perfectionist, i.e., once they deviate from their private opinion they care little about what they exactly state. 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 $\beta \in]0,1[$. 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 4 *Suppose that $P(s)$ is monotonically increasing in the distance from $\bar{s}$, and that 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 mind. 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 fairly agree with the norm speak their minds openly, thus posing mild critique of it. [21]

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 4 induce such properties of P. In the upcoming analysis we will, with some abuse of notation, use y to the denote the distance between the norm and the type who is indifferent between speaking her mind and stating the norm.

Lemma 5 *Suppose there exist some* $\bar{s} \in [-1,1]$ *and* $y > 0$ *such that all types with* $|t - \bar{s}| < y$ *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* $|s - \bar{s}|$ *as long as* $y \leq y_{\max}(\bar{s})$*.*[22]

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 $\bar{s}$.[23]

Put together, Lemmas 4 and 5 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 3 *Suppose* D *is according to (14) and* p *is according to (6) with* $\beta < 1$*. 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* $K \geq K_{\min}(|\bar{s}|)$*.*

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

[22] Note that types are still bounded to the range $[-1,1]$, so those who speak their minds are the types with $t \in [\bar{s} - y, \bar{s} + y] \cap [-1,1]$.

[23] 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 skewed 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 $K_{\min}(|\bar{s}|)$.[24] 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 1), 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.[25]

The second part of the proposition implies that a skewed 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 skewed 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 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, skewness thereby compensates for weakness of social pressure, making skewed norms more sustainable than central norms.

[24] Note that this value is not necessarily equal to the $K_{\min}(|\bar{s}|)$ of Proposition 1.

[25] 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.

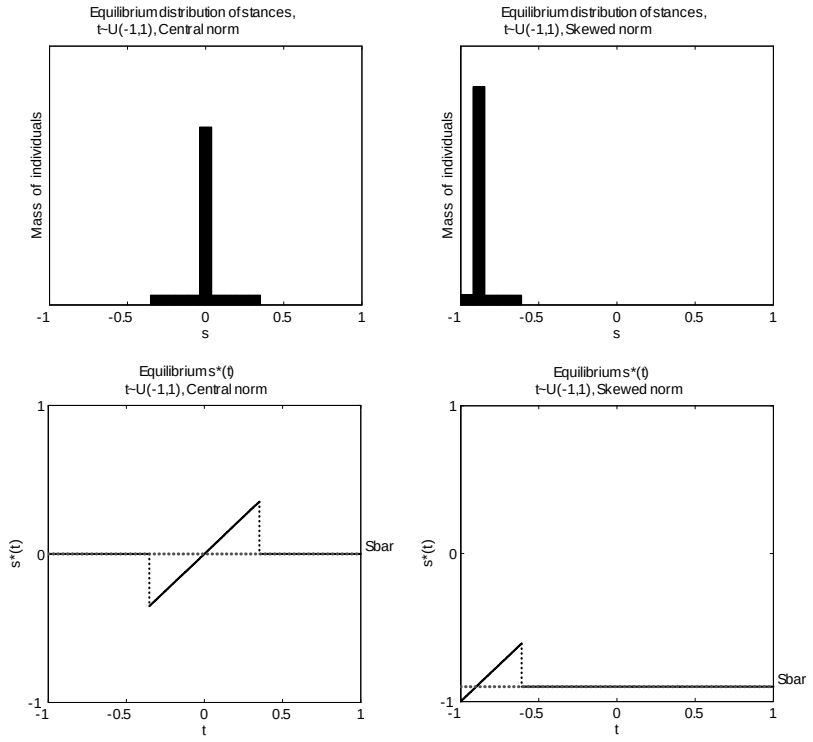

Figure 3: The left graphs show the distribution of stances (top) and $s^*(t)$ in equilibrium (bottom) with a central norm ($\bar{s} = 0$). The right graphs show the distribution of stances and $s^*(t)$ in equilibrium with a skewed norm ($\bar{s} = 0.9$). In all figures $\beta = 0.6$, $\alpha = 0.1$ and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

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. Obviously, the dissonance from deviation determines how reluctant an individual on the right will be to state a different opinion than her bliss point. But 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 preferences of 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, since 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 (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 interesting 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 4 and 5. Lemma 4 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 minimum point (Lemma 5), which will again imply inversion by Lemma 4. 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 4 *Consider the dynamic model in (13) with D being a step function as in (14) and p as given in (6) with $0 < \beta < 1$. 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.*

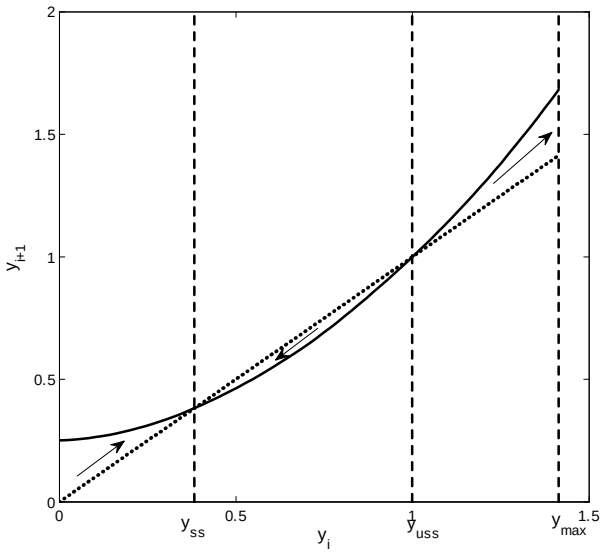

Figure 4: A phase diagram showing convergence to a stable single norm equilibrium when $\bar{s} = -1$, for D being a step function, $\beta = 0.5$ and $K = 2$. The dotted line depicts the diagonal where $y_{i+1} = y_i$, the solid line depicts the intertemporal dynamics $y_{i+1} = f(y_i)$. The vertical lines depict the sufficient conditions for convergence, y_{uss} and $y_{\max}$. 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 3.[26] It may be interesting to note that the steady state is never

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

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 in skewness, 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 skewness. Together, these two observations imply that the cohesiveness of stated opinions in the stable steady state increases with the skewness of the norm.[27]

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, a sufficient condition for convergence is that $y_i < y_{\max}$ and that $y_i < y_{uss}$ whenever it exists.[28] The last point of the proposition says that the sufficient condition for convergence, $y_{conv} \equiv \min\{y_{uss}, y_{\max}\}$, increases with skewness. 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 in skewness. 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 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 skewness.

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

[27] This is another manifestation of the contribution of skewness to the sustainability of single norm equilibria. The intuition here is the same as for why skewed norms enable a lower $K_{\min}$. Once there are fewer individuals on one side (due to skewness 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 extreme opinions expressed in public become more moderate among leftists and rightists alike.

[28] 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.

true even if this group was rather small,[29] 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 equilibrium and still converge back to a steady state with a single norm.[30] 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 skewed the norm is.[31]

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, skewed 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 societies where individual pressure is very concave is that now it is those who fairly agree with the norm who speak their minds, implying that the critique expressed publicly in society will be rather mild.

6 Revolutions and mass protests

Let us now suggest a slightly different interpretation of the model by letting it represent political pressures in society and $\bar{s}$ as a political regime. Essentially, our model provides a microeconomic structure to the models of Granovetter (1978) and Kuran (1989a), henceforth G&K, which analyze social protests, riots and revolutions. In their models, individuals have the binary 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. 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 model then shows the conditions for the existence of a political regime or cluster in equilibrium. In particular, it shows that privately disliked (i.e., skewed or 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. If all individuals are making

[29] I.e., of size $x_0 < x_{ss}$, with $y_0 \in (y_{ss}, y_{conv})$.

[30] 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.

[31] 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.

different statements in equilibrium, we say that no political regime exists, although there will still be political pressure arising from all individual statements. In the appendix we show that if all speak their minds, then a mainstream opinion that minimizes political pressure exists and equals the average private opinion in society.

Note that this result is more elaborate than that of G&K. Our model yields the equilibrium of G&K as one possible case, when society is of the alienating type. Here those who like the regime the most follow it, while those who dislike it the most speak their minds against it. In that case, skewed regimes may exist, but are weaker and essentially require a very high degree of initial conformity in order to exist. The second case, where society is of the inverting type, shows very different properties than that of G&K. Here the political regime is upheld by those who dislike it the most, and those who do speak out against it do it in a very mild manner and are privately rather content with it. 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. In the final case, where individual political pressure is convex, no publicly supported political regimes can exist in the first place.

The dynamic analysis in Propositions 2 and 4 can be used to discuss what may spark the undermining of political regimes in the two main types of societies we discussed, and how the ensuing process will look like. First of all, it is clear that if people start caring less about the political pressure or start valuing their own individual tastes more (i.e., K falls), the political regime may collapse in both types of societies. But the more interesting insights come from considering shifts of opinions and analyzing which individuals will be the ones starting protests against the regime.

We begin our analysis by focusing on alienating societies. 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 skewed 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. This may be a point of no return for the regime. Even if opinions cease to shift further, the

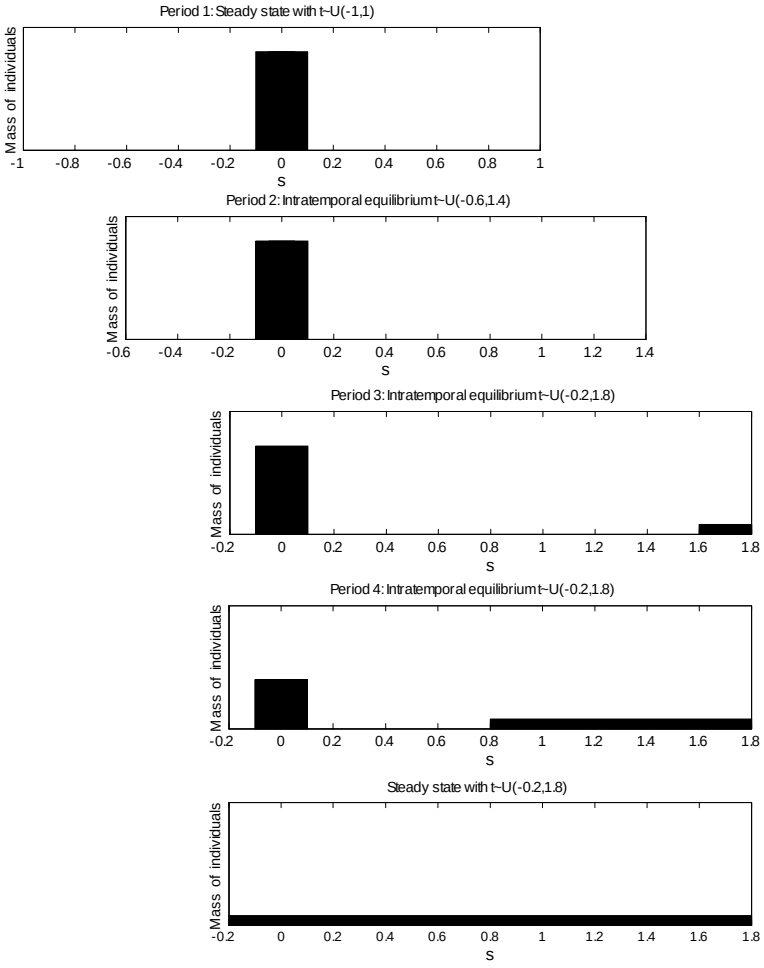

Figure 5: The distribution of stances over time when private sentiments change. $K = 1.3$, $\beta = 0.01$ and $\alpha = 0.5$. The width and placement of each horizontal axis represents the distribution of types in that period. Note that the bar representing the norm at $\bar{s} = 0$ should, strictly speaking, be infinitely narrow. But for clarity of exposition we depict is 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.

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. It will then become observable that the initial regime was actually no longer representative of the private opinions (bottom schedule).

G&K are silent about the post collapse state, since when the individual choice is binary – support the regime or oppose it – one cannot determine whether a successful revolution will lead to a new regime (which may also raise a new opposition) or to a state of pluralism, where each person states her private opinion. In contrast, our model provides a clear prediction – society will converge to pluralism. Moreover, this state will be absorbing, as further changes to the 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 well be established. The overall pattern of revolutions in alienating societies, as described above, seems to provide a reasonable description of the Iranian revolution in 1978-79, which was initiated by the hardest opponents of the old regime but then gained mass support by recruiting more moderate individuals (Razi, 1987). It may also represent the Russian revolution in 1917 whereby the socialists toppled the Tsar.

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 skewed 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.[32] This is through the direct effect of there existing less people on the left and the indirect effect whereby fewer leftists induces more conformity on the right as well. This 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 thing 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

[32]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.

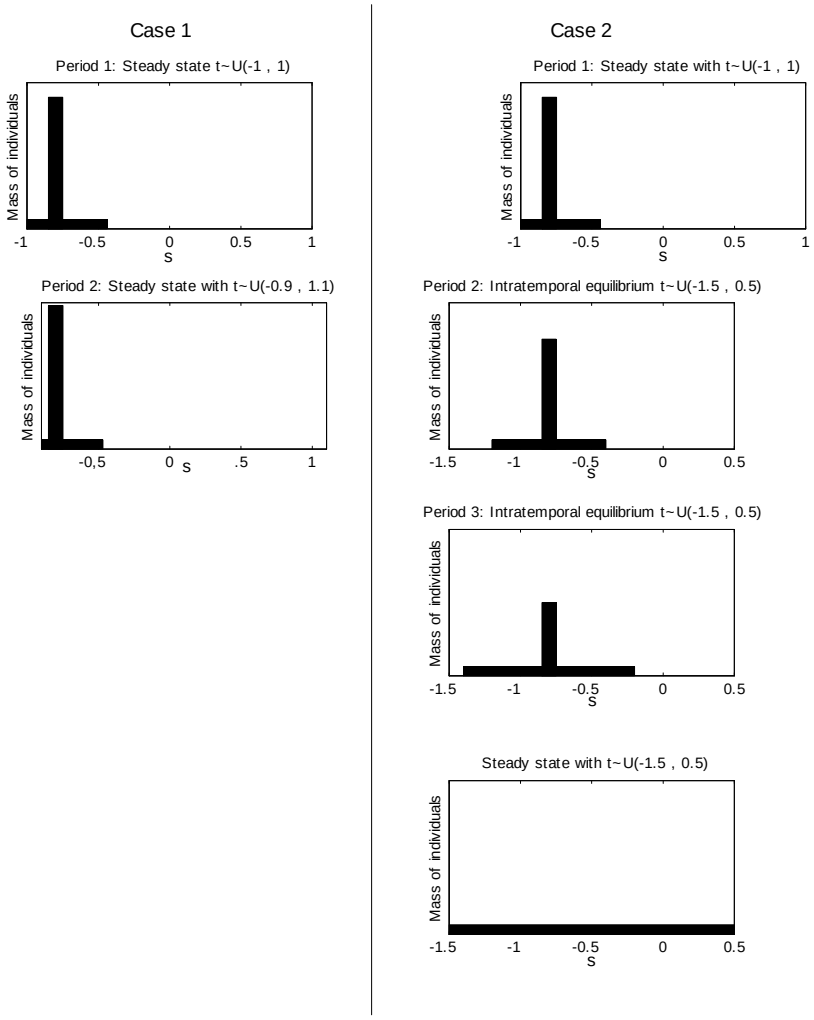

Figure 6: The distribution of stances over time when private sentiments change. $K = 2.1$, $\beta = .5$ and $\alpha = 0.01$. The width and placement of each horizontal axis represents the distribution of types in that period. Note that the bar representing the norm at $\bar{s}$ should, strictly speaking, be infinitely narrow. But for clarity of exposition we depict is 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 has unchanged, but the political regime has collapsed and a new stable steady state has been reached, where are all types speak their minds.

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 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). If the distribution of types changes further, K may become too small to uphold 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 shifts further there will be an equivalent shift in stated opinions. In order to rebuild a regime from this state it is necessary that a group of individuals (or one powerful individual) together dictate a political regime.

The fundamental difference from G&K and the dynamics of the alienating society is that now regimes will break, not through a revolution and fierce critique, but rather following a process of mild critique that evolves from both sides of the regime. 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. This seems to be a reasonable description of the protest movements that led to the collapse 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, the most extreme factions in Egypt (i.e., the Muslim Brotherhood and the Salafi movement) were hardly present in the protests initially (see more below).

Another striking difference from G&K's modeling is that now the undermining of the regime is carried out 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 (though it will eventually ignite opposition also at 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 skewed that there are no opinions on 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.[33] This was a clear pattern in the Arab spring revolution in

[33] Granovetter (1978) points out that in his model protesters may be strange bedfellows in the

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 in Figure 6), the later elections showed that in fact Egyptian society was more conservative than Mubarak's regime. This pattern is exactly what the model predicts (compare the second and fourth schedules from top right, where Mubarak's regime is represented by $\bar{s} = -0.8$).

Now, what 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 with a uniform distribution of types this may happen (under certain parameter settings) if the political regime is skewed so that support is only gained from one side. For instance, in an alienating society with a skewed political regime, the initial protesters are always on one side only. In some circumstances (when $\alpha > 0.5$) this one sided protesting may not gain the sufficient momentum needed to eventually induce protests on the other side too, as is required for a complete collapse.[34] However, if we relax the assumption of a uniform distribution of types, failed revolutions are easier to construct theoretically. Basically, the mechanism would resemble that of a failed revolution in the binary model of G&K. 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 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 additional protesters are supposed to be recruited from this range, the acceleration will slow down. Society then converges to a steady state where many individuals still support the regime, and the revolution fails.

7 Conclusions

This paper studies the sustainability and dynamic stability of social norms and political regimes in the presence of peer pressure. As opposed to settings in which a clear norm exists, as when considering work effort or prosocial behavior, situations that are characterized by peer pressure do not necessarily have a clear norm. Nevertheless, we show that in these situations a clear norm may endogenously evolve,

sense that they have different opinions. But in his model all protesters want to pull the regime in the same direction.

[34] This can be seen in equation (12), which is less steep when y is large. The result described here also requires that the norm is not too skewed, as otherwise there are sufficiently many individuals far away on one side to substantially weaken the norm without the help of individuals from the other side. In an inverting society there is a similar effect. Initially there is protesting on both sides, but if the norm is skewed, say, to the left, then soon enough it will not be possible to recruit more protesters on the left. In this case the additional recruiting will come only from the right, essentially slowing down the acceleration of protests.

be sustainable and often also be dynamically stable. This happens even if individual preferences are heterogeneous. Moreover, it will often be the case that a norm that is skewed with respect to individual preferences will be more sustainable than a representative norm. This can shed light on the sustainability of skewed norms, as observed in religious communities, racial attitudes, honor cultures and various autocratic societies.

Focusing on societies that can give rise to endogenous sustainable norms, the paper highlights a fundamental difference between two main types of societies. In societies where pressure is very concave, types 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 centrally located and to closely represent the opinion of most individuals in society. If society is either very heterogenous, or the norm is skewed, no common norm can be sustained without strong pressure. In the other type of society, where pressure is relatively less 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 dislike the norm 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 skewed norms are more sustainable and more magnetic than central norms.

The model can also be interpreted as being about the formation of political regimes. Naturally, if we would assume the existence of a group with aligned interests, this could lead to additional clustering of opinions by people beyond that group. But what we show is that a regime can be strengthened and 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 that leads 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, like is observed in many revolutions. 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, 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 critique, which will be initially mild. The further evolution will then be a gradual stretching of the freedom of speech.

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 the norm would not exist unless (sufficiently many) individuals actually followed it. But even in situations where a norm *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 this paper has focused on, is not just a technical possibility, but that outcomes will tend to gravitate towards this kind of equilibrium from a broad set of initial conditions.

References

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

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

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

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

[5] 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.

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

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

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

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

[10] Blattman, C., & Miguel, E. (2010). "Civil war". *Journal of Economic Literature*, 48 (1), 3-57.

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

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

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

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

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

[16] 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". *Advances in experimental social psychology*, 24(20), 1-243.

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

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

[19] Cohen, D. (2001). "Cultural variation: considerations and implications". *Psychological bulletin*, 127(4), 451.

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

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

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

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

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

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

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

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

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

[29] Garfinkel, M.R., (1990) "Arming as a Strategic Investment in a Cooperative Equilibrium." *American Economic Review*, 80(1): 50-68.

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

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

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

[33] Goldstone, J. A. (1994). "Is revolution individually rational? Groups and individuals in revolutionary collective action". *Rationality and Society*, 6(1), 139-166.

[34] Granovetter, M., (1976), "Threshold Models of Collective Behavior", *The American Journal of Sociology,* Vol. 83, No. 6, pp. 1420-1443.

[35] Greif, A. (1994). "Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies". *Journal of*

political economy, Vol. 102, No. 5, pp. 912-950.

[36] Hirshleifer, J., (1988). "The Analytics of Continuing Conflict." *Synthèse*, 76(2): 201-33.

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

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

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

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

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

[42] 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.

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

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

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

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

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

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

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

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

[51] 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.

[52] Michaeli, M. & Spiro, D., (2014),"The Distribution of Individual Conformity under Social Pressure across Societies". mimeo, University of Oslo

[53] 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.

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

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

[56] O'gorman, H. J. (1975). "Pluralistic ignorance and white estimates of white support for racial segregation". *Public Opinion Quarterly*, 39(3), 313-330.

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

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

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

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

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

[62] Robinson, C. E. (1932). *Straw votes*. New York: Columbia University Press.

[63] Roland, G. (2004). "Understanding institutional change: fast-moving and slow-moving institutions". *Studies in Comparative International Development*, 38(4), 109-131.

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

[65] 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.

[66] Skaperdas, S., (1992), "Cooperation, Conflict, and Power in the Absence of Property Rights." *American Economic Review*, 82(4): 720-39.

[67] 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.

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

[69] Tarrow S. (1998), *Power in Movement*. New York: Cambridge Univ. Press. 2nd ed.

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

[71] Vandello, J., & Cohen, D. (2000). "Endorsing, enforcing, or distorting? How southern norms about violence are perpetuated". Unpublished manuscript, Princeton University, Princeton, NJ.

[72] 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 & Oswald, 1998) or of work effort (e.g. Kandel & 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, since 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 5

1. *In the single norm equilibria described in Propositions 1 and 3, 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 1), the pressure is reduced only where there is bunching, i.e., at the descriptive norm; and when D is a step function (Proposition 3), 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.[35]

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.[36] 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 skewed – 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 like it fairly much privately. This is since

[35] 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).

[36] 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.

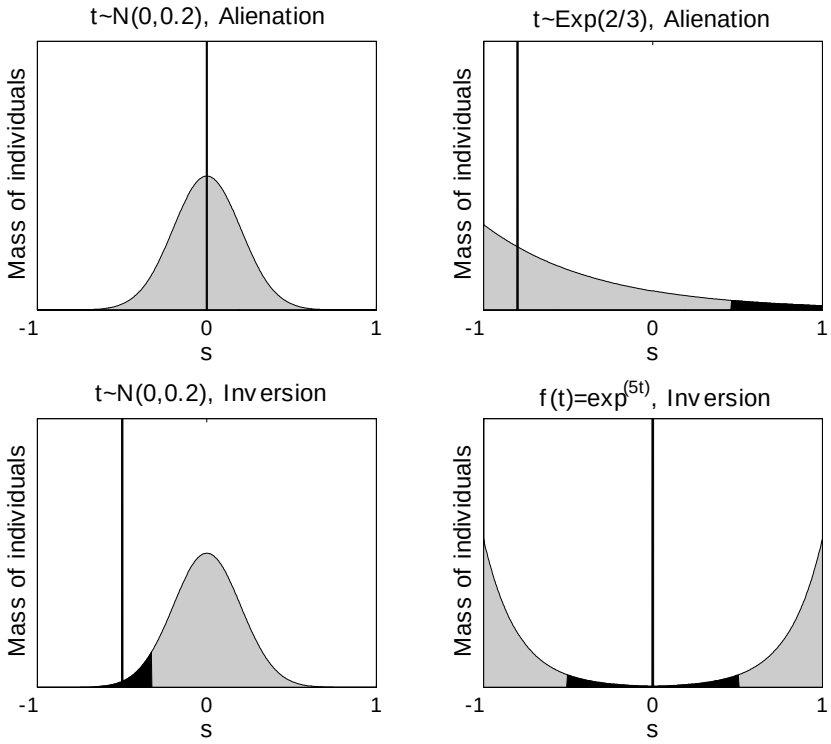

Figure 7: Histograms with single norm steady states in the dynamic model. The black surfaces represent the steady state distribution of stances while the grey surfaces represent the underlying distribution of types. Stance distribution in the zeroth generation is such that all state the same opinion. 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: $\alpha = 0.01$, $\beta = 0.5$, $K = 1.5$, $\bar{s} = 0$.

otherwise there would be a large portion of opposers to the norm, who, by opposing, would make conforming 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 any of the peaks.[37]

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 & Sandholm, 2008) or by an endogenous decision made by the individuals in the previous generation (as, e.g., in Bisin & 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

[37] 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.

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, since 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 & 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 since 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.[38]

As for the dynamic endogenous preference structure of Bisin & 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, since this would fundamentally change the decision problem of the parent.

C Appendix: Proofs (for online publication)

C.1 Initial results

Lemma 6 *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.*

[38] 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 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.

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 Lemma 1

First we note that when $\beta>1$ then p and p' are continuous everywhere. This implies that $P=\int p$ and P' must be continuous everywhere. In particular at $s=\bar{s}$. Hence, $P'|_{s=\bar{s}}$is well defined, and so either $P'|_{s=\bar{s}}=0$ or $P'|_{s=\bar{s}}\neq 0$.

If $P'|_{s=\bar{s}}=0$, 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 $P'|_{s=\bar{s}}\neq 0$ then either $P'|_{s=\bar{s}}>0$ or $P'|_{s=\bar{s}}<0$. If $P'|_{s=\bar{s}}>0$, 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 $|D'(\bar{s};t)|=|P'(\bar{s})|$ because D is strictly concave, and so only this one type can have a local min point of L at $\bar{s}$. This means there will not be a positive mass at $\bar{s}$, which violates the definition of a norm. The same argument applies when $P'|_{s=\bar{s}}<0$.■

C.2 Alienating societies

C.2.1 Proof of Lemma 2

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

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

Since $L(s;t;S)$ is increasing in $|s-t|$ while $P(s)<K$ only if $s=\bar{s}$, 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 3

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 skewed, 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 is $x = (y + 1 - |\bar{s}|)/2$. When $y > 1 + |\bar{s}|$ all types declare the norm.

C.2.3 Proof of proposition 1

Part (1): We will assume that a single norm equilibrium exists at $|\bar{s}|$ and prove that the assumption holds if and only if $K \geq K_{\min}(|\bar{s}|)$. Since Lemma 2 implies that, given a single norm, $s^*(t)$ 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 3 would equal the value of x that was initially assumed in Lemma 2 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

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

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. 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

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

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 $x_i=\left(1-|\bar{s}|\right)^{\alpha}/K$. 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.[39] Substituting $x_i=\left(1-|\bar{s}|\right)^{\alpha}/K$ in equation (18) yields $G=\left(1-|\bar{s}|\right)-\left(1-|\bar{s}|\right)^{\alpha}/K$, which equals 0 when $K=\left(1-|\bar{s}|\right)^{\alpha-1}$. 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 $K=\left(1-|\bar{s}|\right)^{\alpha-1}$. 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

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

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

Part (2) follows directly from the fact that $\left(1-|\bar{s}|\right)^{\alpha-1}$ and $\left(1+|\bar{s}|\right)^{\alpha}$ are both increasing in $|\bar{s}|$.■

C.2.4 Proof of proposition 2

The proof of the proposition builds on a few preliminary results and auxiliary lemmas, which will be presented first.

Note first that Lemmas 2 and 3 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 $x_{i+1}=f\left(x_i;K,|\bar{s}|\right)$ is convex within each of the first two regions and has a kink at the border between the regions. Together with $G'\left(0;K,|\bar{s}|\right)<0$ (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.

[39] 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 $G\left(1,K,|\bar{s}|\right)<0\Rightarrow G\left(x_i,K,|\bar{s}|\right)<0 \ \ \forall x_i\in\left]0,1\right]$, hence no single norm equilibrium exists (we know that $G\left(1,K,|\bar{s}|\right)<0$ because $K^{1/\alpha}<1+|\bar{s}|$).

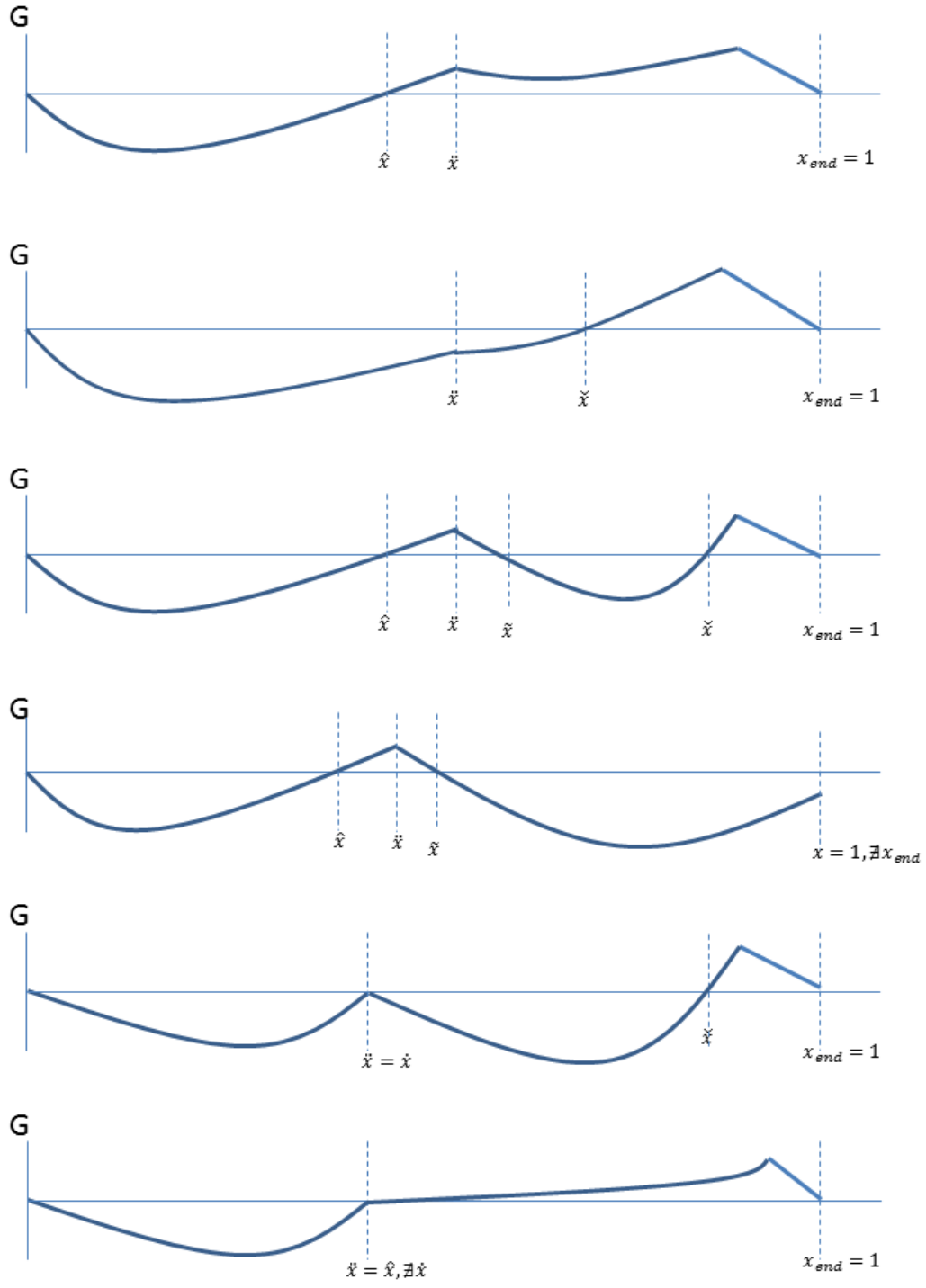

Figure 8: Some variations of the G function of equation (18), depicting the potential fix points defined in equation (22). Note that these variations of G are not exhaustive but are only meant to complement the proof.

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

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 7 *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 8 *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 if $|\bar{s}|$ increases, region (2) expands at the expense of region (1). Then, by Lemma 7, we get that G' is weakly falling in $|\bar{s}|$.* ■

Lemma 9 *1) If $\hat{x}$ exists then it is independent of $|\bar{s}|$. 2) If $\breve{x}$ 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 $\breve{x}$ is in region 2. Lemma 8 together with $G(0) = 0$ imply that G is weakly falling in $|\bar{s}|$ in region 1 and 2. Combined with the fact that $G'(\breve{x}) > 0$ (by definition) this implies $\breve{x}$ (if it exists) is weakly increasing in $|\bar{s}|$. 3) Same logic as part 2 but now with $G'(\tilde{x}) < 0$.* ■

Lemma 10 *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 11 *Suppose $K > K_{\min}$. Let $x_{conv} \equiv \{x_i : x_i = \min\{\hat{x}, \breve{x}\}\}$ (when $\hat{x}$ or $\breve{x}$ 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.*

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.*[40] *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 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* $x_{i+1} < x_i \quad \forall x_i < x_{conv}$*. Thus, the instability of* x_{conv} *implies that* $x_{ss}(|\bar{s}|) = 0$*.*

1) In the proof of Proposition 1 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* $\dot{x}$*. This concludes the second subsentence.*

4) $x_{conv} \equiv \hat{x}$ *whenever* $\exists \hat{x}$*. From Lemma 9 we know that* $\hat{x}$ *is independent of* $|\bar{s}|$ *and from Lemma 10 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 9 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* $H = (\check{x}K)^{1/\alpha} +$

[40] 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 $x_{end} = \hat{x}$.

$1-|\bar{s}|-2\check{x}=0$ *defining* $\check{x}$*. Using the implicit function theorem we get* $d\check{x}/dK = -\left(\check{x}\right)^{1/\alpha}K^{1/\alpha-1}/\alpha/\left(K^{1/\alpha}\left(\check{x}\right)^{1/\alpha-1}/\alpha-2\right)<0\Leftrightarrow K^{1/\alpha}\left(\check{x}\right)^{1/\alpha-1}>2\alpha$*. 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 can never switch from* $x_{conv}=\hat{x}$ *to* $x_{conv}=\check{x}$*. This implies that* x_{conv} *is decreasing in* K *also globally.* ■

Lemma 12 *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 1). 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 9 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 2

Part 1: The 'if' part follows from Lemma 12. As for the 'only if' part, we showed in the proof of Proposition 1 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 12.

Part 3: Follows from Lemma 11.

Part 4: Follows from Lemma 11.■

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 4

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 5

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 6 regarding the contribution from the uniform part). The first statement of the lemma then follows. 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 $P'(s)$

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

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

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

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 6 implies that $P(s)$ is increasing in the distance from $0 > \frac{-1+\bar{s}+y}{2}$. Differentiating with respect to y we get

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

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 $y \leq y_{\max}(\bar{s})$.[41] 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

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

Differentiating with respect to s_l we get

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

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

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

we get[42]

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

and so, if $\beta < 1$

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

and if $\beta = 1$

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

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

[41] 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.

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

C.3.3 Proof of Proposition 3

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 13 *Suppose that $\beta \leq 1$ and $\bar{s} \in [-1, 0]$. Denote the value of y that solves the equation $\left(1 - \frac{y}{2} - \frac{s_l}{2}\right) \beta y^{\beta-1} = y^\beta - \frac{1}{2}(s_l + y)^\beta$ by $y_{\min}(\bar{s})$. Then $y_{\min} \leq 1$.*
Proof. *(i) For a given $\bar{s}$, we can simply write $y_{\min}(\bar{s}) = y_{\min}$. We then have*

$$\begin{aligned}
\left(1 - \frac{y_{\min}}{2} - \frac{s_l}{2}\right) \beta y_{\min}^{\beta-1} &= y_{\min}^\beta - \frac{1}{2}(s_l + y_{\min})^\beta \Rightarrow \\
(1 - y_{\min} - \bar{s}) \beta y_{\min}^{\beta-1} &= 2y_{\min}^\beta - (\bar{s} + y_{\min} + 1)^\beta \\
\frac{dLHS}{dy} &= -\beta y_{\min}^{\beta-1} + (\beta - 1)(1 - y_{\min} - \bar{s}) \beta y_{\min}^{\beta-2} < 0 \\
\frac{dRHS}{dy} &= 2\beta y_{\min}^{\beta-1} - \beta(\bar{s} + y_{\min} + 1)^{\beta-1} > 0
\end{aligned}$$

Hence, there is only one intersection, a unique y. At $y = 1$ we have

$$\begin{aligned}
LHS|_{y=1} &= -\bar{s}\beta \\
RHS|_{y=1} &= 2 - (\bar{s} + 2)^\beta,
\end{aligned}$$

and since $\beta \leq 1$ and $\bar{s} \in [-1, 0]$ we have $(2 + \bar{s})^\beta \leq 2 + \bar{s} \leq 2 + \bar{s}\beta$, and so $LHS|_{y=1} \leq RHS|_{y=1}$, which implies that $LHS < RHS$ for any $y > 1$. Hence, $y_{\min} \leq 1$ for any $\bar{s}$ and β. ■

Lemma 14 *$K_{\min}$ is weakly decreasing in $|\bar{s}|$.*
Proof. *First note that $K_{\min}$ is never found on the border between the regions (1) and (2), 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 (34) as a function of $\bar{s}$ for the two regions and differentiate $1/K$ w.r.t. $\bar{s}$. This yields*

$$\frac{d(1/K)}{d\bar{s}} = \begin{cases} 0 \text{ if } y \leq \bar{s} + 1 \\ -\frac{y^\beta}{2} + \frac{1}{2}(\bar{s} + y + 1)^\beta - \frac{1}{2}(\bar{s} + 1)^\beta \text{ if } y \in [\bar{s} + 1, \min\{y_{\max}(\bar{s}), 1 - \bar{s}\}] \end{cases} \tag{29}$$

$$\frac{d^2(1/K)}{d\bar{s}^2} = \begin{cases} 0 \text{ if } y \leq \bar{s} + 1 \\ \frac{1}{2}\beta(\bar{s} + y + 1)^{\beta-1} - \frac{1}{2}\beta(\bar{s} + 1)^{\beta-1} \text{ if } y \in [\bar{s} + 1, \min\{y_{\max}(\bar{s}), 1 - \bar{s}\}] \end{cases} \tag{30}$$

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 $\frac{d^2(1/K)}{d\bar{s}^2} \leq 0$ when $\beta < 1$. 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 (34)). Hence, since we have been analyzing the case of $\bar{s} \leq 0$, more generally K 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 $|\bar{s}|$. ■

Lemma 15 *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 while $s^*(t) = \bar{s}$ for the rest. If type $t = \bar{s}+y$ is indifferent between the two corner solutions $s^*(t) = \bar{s}$ and $s^*(t) = t$, then for all types we have*

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

Proof. *For types with $t > \bar{s}$ the result follows from Lemmas 4 and 5. 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, 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 5 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 16 *Let $\bar{s} \in [-1,1]$ and let D be given by (14), and suppose that $\beta < 1$. 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 that indeed $s^*(t) = \bar{s}$ for all types with $|t-\bar{s}| > y$ and $s^*(t) = t$ for all types with $|t-\bar{s}| < y$. Then $K(y)$ has either a U-shape or a W-shape.*

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 5 that P is increasing in $|\sigma|$ (recall $\sigma \equiv s - \bar{s}$). Moreover, from Lemma 4 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 15 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* $y > s_l$*). 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:*

$$P(\sigma)=\begin{cases} Kx|\sigma|^{\beta}+K\frac{1}{2}\frac{(|\sigma|+y)^{\beta+1}+(y-|\sigma|)^{\beta+1}}{\beta+1} & \text{if } |\sigma|\leq y \\ Kx|\sigma|^{\beta}+K\frac{1}{2}\frac{(|\sigma|+y)^{\beta+1}-(|\sigma|-y)^{\beta+1}}{\beta+1} & \text{if } |\sigma|>y \end{cases} \tag{31}$$

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

$$\begin{aligned} 1/K+\frac{1}{2}\frac{2y^{\beta+1}}{\beta+1} &= (1-y)y^{\beta}+\frac{1}{2}\frac{(2y)^{\beta+1}}{\beta+1} \\ &\Rightarrow 1/K=(1-y)y^{\beta}+\left(2^{\beta}-1\right)\frac{y^{\beta+1}}{\beta+1}. \end{aligned} \tag{32}$$

Region (2): $y > s_l$

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*

$$\begin{aligned} 1/K+\frac{1}{2}\frac{(s_l)^{\beta+1}+(y)^{\beta+1}}{\beta+1} &= \left(1-\frac{y}{2}-\frac{s_l}{2}\right)y^{\beta}+\frac{1}{2}\frac{(s_l+y)^{\beta+1}}{\beta+1} \Rightarrow \\ 1/K &= \left(1-\frac{y}{2}-\frac{s_l}{2}\right)y^{\beta}+\frac{1}{2}\frac{(s_l+y)^{\beta+1}-(s_l)^{\beta+1}-y^{\beta+1}}{\beta+1} \end{aligned} \tag{33}$$

Joining the two regions:

Following equations 32 and 33, we can get the following expression for $\frac{1}{K}(y)$.

$$\frac{1}{K}(y)=\begin{cases} (1-y)y^{\beta}+\left(2^{\beta}-1\right)\frac{y^{\beta+1}}{\beta+1} & \text{if } y\leq s_l \\ \left(1-\frac{y}{2}-\frac{s_l}{2}\right)y^{\beta}+\frac{1}{2}\frac{(s_l+y)^{\beta+1}-(s_l)^{\beta+1}-y^{\beta+1}}{\beta+1} & \text{if } y\in[s_l,\min\{y_{\max}(\bar{s}),2-s_l\}] \end{cases} \tag{34}$$

Differentiating in both regions yields

$$\frac{d(1/K)}{dy}=\begin{cases}(1-y)\,y^{\beta-1}\beta-y^{\beta}\left(2-2^{\beta}\right) & \text{if } y\leq s_l\\ \left(1-\frac{y}{2}-\frac{s_l}{2}\right)\beta y^{\beta-1}-y^{\beta}+\frac{1}{2}\left(s_l+y\right)^{\beta} & \text{if } y\in\left[s_l,\min\left\{y_{\max}\left(\bar{s}\right),2-s_l\right\}\right]\end{cases} \tag{35}$$

and

$$\begin{aligned}&\frac{d^2(1/K)}{dy^2}\\ &=\begin{cases}-y^{\beta}\beta+(1-y)\,y^{\beta-2}\left(\beta-1\right)\beta-\beta y^{\beta-1}\left(2-2^{\beta}\right)<0 & \text{if } y\leq s_l\\ \left(1-\frac{y}{2}-\frac{s_l}{2}\right)\beta\left(\beta-1\right)y^{\beta-2}-\frac{3}{2}\beta y^{\beta-1}+\frac{1}{2}\beta\left(s_l+y\right)^{\beta-1}<0 & \text{if } y\in\left[s_l,\min\left\{y_{\max}\left(\bar{s}\right),2-s_l\right\}\right]\end{cases}\end{aligned} \tag{36}$$

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. Differentiating at that point at each region yields

$$\frac{d\left(1/K\right)}{dy}|_{y\rightarrow^{-}s_l}=\left(s_l\right)^{\beta-1}\left[\beta+\left(2^{\beta}-\beta-2\right)\left(s_l\right)\right]\quad \textit{and}$$

$$\frac{d\left(1/K\right)}{dy}|_{y\rightarrow^{+}s_l}=\left(1-s_l\right)\beta\left(s_l\right)^{\beta-1}-\left(s_l\right)^{\beta}+\frac{1}{2}2^{\beta}\left(s_l\right)^{\beta}=\left(s_l\right)^{\beta-1}\left[\beta+\left(2^{\beta-1}-\beta-1\right)\left(s_l\right)\right].$$

Noting that $2^{\beta-1}-1<0$, we get that $2(2^{\beta-1})-\beta-2\leq 2^{\beta-1}-\beta-1$, and so $\frac{d(1/K)}{dy}|_{y\rightarrow^{+}s_l}$ is greater (strictly, unless $s_l=0$) than $\frac{d(1/K)}{dy}|_{y\rightarrow^{-}s_l}$.

Next, note that in the first region, where $y\leq s_l$, we get the following.

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

With equation 36, this implies that for a large enough value of s_l (so that y can approach 1 while in the first region), $\frac{1}{K}(y)$ gets a local maximum in the first region. The concavity of $\frac{1}{K}(y)$ in both regions, together with the increase in $\frac{d(1/K)}{dy}$ at the border between the regions, imply that $\frac{1}{K}(y)$ will be either hill-shaped or M-shaped, and so $K(y)$ will be either U-shaped or W-shaped.

Otherwise, if $\frac{1}{K}(y)$ is only increasing in region (1), then $\frac{d(1/K)}{dy}|_{y\rightarrow^{+}s_l}\geq\frac{d(1/K)}{dy}|_{y\rightarrow^{-}s_l}>0$, implying that $\frac{1}{K}(y)$ will have a max point in region (2) if the FOC holds in this region. Using equation (35), we get the following FOC in region (2):

$$\frac{d\left(1/K\right)}{dy}=\left(1-\frac{y}{2}-\frac{s_l}{2}\right)\beta y^{\beta-1}-y^{\beta}+\frac{1}{2}\left(s_l+y\right)^{\beta}=0$$

Region (2) is bounded from below by $y=s_l$, and from above by $\min\left\{y_{\max}\left(\bar{s}\right),2-s_l\right\}$ ($y\leq y_{\max}\left(\bar{s}\right)$ by assumption, and $y\leq 2-s_l$ by construction). If $y_{\max}\left(\bar{s}\right)\geq 2-s_l$, then the fact that

$$\frac{d\left(1/K\right)}{dy}|_{y=2-s_l}=-\frac{1}{2}\left(2-s_l\right)^{\beta}+\frac{1}{2}\left(2^{\beta}-\left(2-s_l\right)^{\beta}\right)=2^{\beta-1}-\left(2-s_l\right)^{\beta}<2^{\beta-1}-1<0$$

implies that the FOC holds in region (2). Otherwise $y_{\max}(\bar{s}) < 2 - s_l$. *In this case, Lemma 13 shows that the value of* y *that solves the FOC is always smaller than* 1. *So the fact that* $1 \leq y_{\max}(\bar{s})$ *implies once again that the FOC holds in region (2). Hence, we get in this case that* $\frac{1}{K}(y)$ *is hill-shaped implying* $K(y)$ *is U-shaped.* ■

Proof of Proposition 3

From Lemma 16 we know that $K(y)$, the function that describes the pairs (K, y) for which a single norm equilibrium exists, is either U-shaped or W-shaped. 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 $K_{\min}$ denote the minimal value of $K(y)$. It thus immediately follows that for $K \geq K_{\min}$ there exists a fix point y while for $K < K_{\min}$ there does not. The fact that $K_{\min}$ is weakly decreasing in $|\bar{s}|$ follows directly from Lemma 14.■

C.3.4 Proof of Proposition 4

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 17 *Suppose in some generation* i *there exists a cutoff distance from the norm, such that all types in that generation that fulfill* $|t - \bar{s}| > y_i$ *declare the norm and all types that fulfill* $|t - \bar{s}| \leq y_i$ *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 5* 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 15 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:

$$F= \tag{37}$$

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

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

$$F_{y_{i+1}}=-\frac{dP}{dy_{i+1}}=-\frac{dP}{ds}|_{s=y_{i+1}}, \tag{38}$$

which we know is negative by Lemma 5. 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 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 $\frac{1}{2}y_i^{\beta}+\frac{1}{2}y_{i+1}^{\beta}>\frac{1}{2}(y_{i+1}+y_i)^{\beta}$ and $\frac{1}{2}y_i^{\beta}>\frac{1}{2}(y_i-y_{i+1})^{\beta}$. ■

Lemma 18

1. *$y_{\max}(\bar{s})$ (from Lemma 5) is increasing in $|\bar{s}|$.*
2. *Let $\tilde{y}$ denote a solution to the equation (33). If $K'(\tilde{y})>0$, $\tilde{y}$ is increasing in $|\bar{s}|$, and if $K'(\tilde{y})<0$, $\tilde{y}$ is 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 5 we show that it is unique for a given $\bar{s}$ such that $P(s)$ is monotonically increasing if and only if $y \leq y_{\max}(\bar{s})$. 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 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 $y \leq y_{\max}(s_l)$, i.e., $y_{\max}(s_l)$ has decreased. This concludes the proof of statement (1). 2) Rewriting equation (33) as a function of $\bar{s}$, we get*

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

From the second regions of equations (29) and (30) we have the following derivatives of $1/K$ w.r.t. $\bar{s}$.

$$\frac{d(1/K)}{d\bar{s}} = -\frac{y^{\beta}}{2} + \frac{1}{2}(\bar{s}+y+1)^{\beta} - \frac{1}{2}(\bar{s}+1)^{\beta}$$
$$\frac{d^2(1/K)}{d\bar{s}^2} = \frac{1}{2}\beta(\bar{s}+y+1)^{\beta-1} - \frac{1}{2}\beta(\bar{s}+1)^{\beta-1}$$

Note that $\frac{d(1/K)}{d\bar{s}}|_{y=0} = 0$, $\frac{d^2(1/K)}{d\bar{s}^2}|_{y=0} = 0$ and $\frac{d^2(1/K)}{d\bar{s}^2}|_{y>0} < 0$ when $\beta < 1$. Since we are looking at $\bar{s} < 0$ these results imply that $K(y)$ is monotonically increasing in $\bar{s} < 0$ for any given y, i.e., the whole function of $K(y)$ is monotonically decreasing in $|\bar{s}|$. In Lemma 16 we showed that $K(y)$ is either U-shaped or W-shaped. It thus follows that where $K'(y) > 0$, y is increasing in $|\bar{s}|$, and where $K'(y) < 0$, y is decreasing in $|\bar{s}|$. This holds in particular for $\tilde{y}$. ■

Proof of Proposition 4

1) Recalling that $F = 0$ in equation (37) 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 $y_{i+1}(0) > 0$.[43] Lemma 17 further shows that y_{i+1} is an increasing function of y_i. If $K < K_{\min}(|\bar{s}|)$, we know from Lemma 16 that no steady state exists. Otherwise, if $K \geq K_{\min}(|\bar{s}|)$, then by Lemma 16 we know that a steady state exists (at least one). Next, note that F in equation (37) 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 38) 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.

[43] 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 $F = 0$.

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}, we also know that $y_{ss} \leq y_{\min}(\bar{s})$ (as defined in Lemma 13), because our analysis up till now implies that $y_{i+1}(y_{\min}(\bar{s})) < y_{\min}(\bar{s})$.[44] From $y_{i+1}(0) > 0$ we know that $y_{ss} \neq 0$, and since $y_{ss} \leq y_{\min}(\bar{s})$, 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 used in Lemma 16.[45] If the steady state falls in the first region, where $y < s_l$, then F_4 applies. There we have

$$\begin{aligned}
\frac{dy_{i+1}}{dy_i} &= -\frac{F_{y_i}}{F_{y_{i+1}}} \\
&= -\frac{y_t^\beta + y_{t+1}^\beta - \frac{1}{2}(y_{t+1} + y_t)^\beta - \frac{1}{2}(y_{t+1} - y_t)^\beta}{-\left[(1 - y_t)\beta y_{t+1}^{\beta-1} + \frac{1}{2}(y_{t+1} + y_t)^\beta - \frac{1}{2}(y_{t+1} - y_t)^\beta\right]} \\
&= \frac{2y_i^\beta - 2^{\beta-1}y_i^\beta}{\left[(1 - y_i)\beta y_i^{\beta-1} + 2^{\beta-1}y_i^\beta\right]}
\end{aligned} \tag{39}$$

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 35 to 0 and solve for y). From Lemma 16 we know

[44] 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.

[45] 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.

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 we have

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

which is smaller than 1 iff

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

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 (35). 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 (34) 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 18 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 (i.e., when $y_i \to y_{uss}$ from both sides), if such 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 $y_i \to^- y_{uss}$, $y_{i+1} < y_i$, and so there is convergence to a stable steady state from any $y_i < y_{uss}$.[46]

[46] 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.

4) Revisiting Lemma 18, 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 5

1) In the single norm equilibria in Proposition 1, 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 3, $y \leq y_{\max}$ (see the proof of that proposition). By Lemma 5 we know that P is increasing in the distance from $\bar{s}$ whenever $y \leq y_{\max}$.

2) Follows from Lemma 6.■