# Norm Conformity across Societies*

Moti Michaeli† & Daniel Spiro‡

**Abstract**

This paper studies the aggregate distribution of declared opinions and behavior when heterogeneous individuals make the trade-off between being true to their private opinions and conforming to a social norm. The model sheds light on how various sanctioning regimes induce conformity and by whom, and on phenomena such as societal polarization and unimodal concentration. In strict societies, individuals will tend to either fully conform to the social norm or totally ignore it, while individuals in liberal societies will tend to compromise between these two extremes. Furthermore, the degree of strictness determines whether those who nearly agree with the norm or those who strongly disagree with it will conform. The degree of liberalism similarly determines which individuals will compromise the most. A number of empirical predictions, and several methods of how to test them, are suggested.

Keywords: Social pressure, Conformity, Liberal, Strict.
JEL: D01, D30, D7, K42, Z1, Z12, Z13

*We wish to thank Laurie Anderson, Florian Biermann, Bård Harstad, Sergiu Hart, John Hassler, Arie Kacowicz, Edwin Leuven, Andrea Mattozzi, Karine Nyborg, Andrew Oswald, Ignacio Palacios-Huerta, Alyson Price, Torsten Persson, Francesco Trebbi, Jörgen Weibull, Robert Östling, two anonymous referees and seminar participants at the Hebrew University, the University of Oslo, Stockholm University and the Stockholm School of Economics for valuable comments.

†Corresponding author. The European University Institute, Florence. *E-mail*: motimich@gmail.com. Tel. [+39] 055 4685 901. Postal address: Via della Piazzuola 43 I-50133 Firenze.

‡Department of Economics, University of Oslo, daniel.spiro@econ.uio.no.

# 1 Introduction

It is by now well established that social norms, and social pressure to conform to these norms, influence individual decision making in a wide spectrum of situations. In particular, imagine a controversial social or political issue where there exists a social norm, that is, a consensual opinion or norm of behavior. Suppose now that each individual in society has some private opinion regarding this issue, and each needs to publicly declare her stance. An individual whose private opinion differs from the social norm will need to consider the trade-off between the social pressure of violating the norm and the psychological cost of stating an opinion different to her private view. In many cases, such as at what age to bear children, how much alcohol to drink and to what extent to follow religious customs, the individual can choose the extent of conformity to the norm from a continuum. We analyze this basic trade-off in a heterogeneous agent framework and present the aggregate outcomes across societies.

In particular, we examine the *extent* of conformity that one person exhibits compared to that exhibited by another person with a different private opinion. This analysis provides predictions for (i) which individuals in society will conform more, (ii) which individuals in society will make larger individual concessions, (iii) the distribution of stated opinions in society and (iv) which norms will be sustainable. We show that although the problem faced by each individual is fairly simple, the outcomes at the aggregate level are diverse, and we analyze how these outcomes depend on the underlying characteristics of society.

In practice, societies differ not only in the general weight of social pressure, but also in its curvature. That is, they differ in the way they sanction small deviations from the norm compared to large ones. We show that the curvature of social pressure has more intricate and possibly more important effects than the general weight of pressure. Moreover, in order to connect the model's results to outcomes across societies, and drawing on observations of sanctioning in different societies and cultures (to be presented in the next section), we apply labels to the curvature of social pressure: *strict* societies are those emphasizing full adherence to the social norm, and hence they utilize concave social pressure; *liberal* societies are those allowing freedom of expression as long as it is not too extreme, and hence they utilize convex social pressure. Strictly speaking, these labels are not necessary for the formal analysis, but they prove useful, as they highlight the consistency between the results of the model and observations of actual societies.

We find that in liberal societies, the convexity of social pressure facilitates a compromise mentality, where most individuals are compelled to adjust at least a little bit to the norm. Furthermore, the degree of

liberalism (i.e., the degree of convexity) plays an important role. Very liberal societies will tend mainly to make those who privately detest the norm adjust to it. This will create a society that looks polarized. Less liberal societies will be more directed at getting moderates to conform and hence will look cohesive, with a concentration of stances around the norm.

Strictness, on the other hand, facilitates an all-or-nothing mentality, since only full conformity counts. This may indeed lead to full conformity, but may also backfire so that some individuals do not concede at all. Moreover, the degree of strictness (i.e., the degree of concavity) is important in predicting who follows the norm. In very strict societies, the full conformers are those who nearly agree with the norm anyway, while those who strongly reject the norm privately, express their dissent publicly as well. However, in less strict societies, paradoxically those who dislike the norm the most are the only ones upholding it, while those who basically agree with the norm privately, criticize it mildly in public. This creates a surprising result: an inversion of opinions.

We also find that, in some cases, opposition to the norm will be more extreme in strict societies than in liberal ones. This result is surprising as it emerges even when sanctions are harsher in strict societies. It is driven by the all-or-nothing behavior of individuals in strict societies, compared to the compromising behavior of individuals in liberal ones. This result is formalized into a testable prediction and we suggest some methods and situations of social interaction in which this and a few other predictions can be tested.

Another outcome that clearly separates liberal and strict societies relates to the possible location of the norm. Letting the norm be the average *declared* opinion in society, we show that norms in liberal societies are bound to be representative also of the *private* sentiments in society, as the norm coincides with the average private opinion. In contrast, strict societies may well maintain a biased social norm, centered on a point that is far from the average private opinion. This implies that strict societies allow for multiple equilibria, while liberal societies do not. One interpretation of this result is that strictness is a tool for maintaining biased norms.

The contribution of our paper lies in explaining different patterns of norm conformity across societies. This requires modeling continuous choice under various sanctioning regimes. Previous theoretical papers with a similar individual trade-off usually model binary decisions (e.g., Bénabou & Tirole 2011; Brock & Durlauf 2001; Lindbeck et al. 2003; Lopéz-Pintado & Watts 2008; Akerlof 1980; and Kuran 1995). Models of continuous decisions usually assume quadratic utility functions (Kuran

and Sandholm 2008 and Manski and Mayshar 2003), thus limiting their ability to analyze how the sanctioning regime affects conformity. Another type of model (see Bernheim 1994 and Bénabou and Tirole 2006) assumes an exogenous norm in a signaling game. There individuals are punished or rewarded for their *private* preferences, instead of their declarations or actions as in our model. Finally, our paper is related to the works of Eguia (2013) and Clark & Oswald (1998), who, although analyzing different issues than we do, do concentrate on how the curvature of preferences affects individual behavior.[1]

The next section motivates our labels by considering observations of sanctioning across societies. The model is outlined in section 3. Section 4 presents the main differences between liberal and strict societies and Sections 5 and 6 analyze liberal and strict societies respectively in more detail. Section 7 presents a number of testable model implications and suggests some methods and data sources for carrying out these tests. Section 8 concludes. Proofs are covered in the appendix.

## 2 Social pressure across societies

In this section we demonstrate that an important distinction between societies concerns the relative strength of sanctions they impose on small versus large deviations from the norm. One example comes from experiments using public goods games with punishment (Herrmann et al. 2008). In these games participants punish others who contribute a different amount to a public good than they themselves do. The experimental results suggest that deviations are punished convexly in places such as Copenhagen, Bonn and Melbourne, while they are punished concavely in places such as Riyadh and Muscat. Another detail to note in the results is that for large deviations, heavier punishments were used in Melbourne compared to those used in either Riyadh or Muscat, while for small deviations the opposite applies. This pattern matches that of the stylized societies 2 (representing Muscat and Riyadh) and 3 (representing Melbourne) in Figure 1.

A more anecdotal demonstration of these points emerges from a crude comparison of the sanctioning systems in the Israeli Jewish Ultraorthodox community, or under the Taliban, with those of liberal West European institutions.[2] An important difference between the Taliban and the Ultraorthodox sanctioning systems is that the Taliban use substantially

[1] In a subsequent paper, Michaeli & Spiro (2014), we study the conditions for the very existence of an endogenous social norm when all individuals put pressure on each other.

[2] This is to some extent a comparison of informal and formal sanctioning, but the purpose here is to highlight that sanctioning systems vary in curvature.

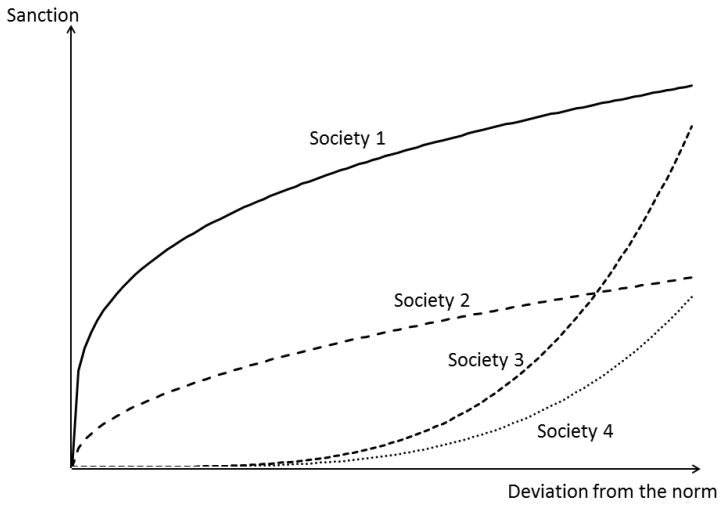


Figure 1: Sanctioning across societies. A system of sanctioning may be at the same time harsh and concave (society 1). Alternatively, it may be light and concave (society 2). Or, it may be harsh and convex (society 3). Finally, like in society 4, it may be light and convex.

heavier sanctions for any comparable deviation from the norm. But one characteristic they have in common is that they require strict adherence to their code of conduct, sanctioning any small deviation harshly, while large deviations are sanctioned only slightly more.[3] Hence, they respectively match stylized societies 1 and 2 in Figure 1.

What about the sanctioning structure of liberal West European institutions? Almost by definition, and as is manifested in their constitutions, liberal democracies allow citizens a broad freedom of expression and political parties a wide range of positions. But once a party or an individual expresses views very far from the consensus, the sanction is ramped up.[4] This suggests that liberal democracies will tend to be convex in how they deal with deviations constitutionally, like societies 3 and 4 in Figure 1.

As incomplete and stylized as these descriptions may be, they do

[3]There are numerous accounts of the Taliban using capital punishment for both misdemeanor and larger offenses. In Israeli Ultraorthodox society, a woman may be censured for wearing a dress that is too short, and a man for publicly supporting the drafting of members of the Ultraorthodox community into the Israeli army.

[4]For example, a party that wants to abolish democracy may become illegal (like Nazi parties are in certain countries). Likewise, an individual who openly expresses extreme right-wing or extreme left-wing opinions, or supports Sharia Law, may be subject to surveillance and in some cases even fined or arrested.

highlight that a sanctioning system should be represented by both its general harshness and its curvature. They also highlight that societies often considered to be strict are the ones using concave sanctioning, thus meticulously punishing minor deviations from the norm but not distinguishing much between large and small wrongdoings. Similarly, liberal societies are those using convex sanctioning, in doing so allowing broad freedom of speech around the norm.

## 3 The model

An individual is represented by a type $t \in [t_l, t_h] \subset \mathbb{R}$, which is a point on an axis of opinions. Let $s$ be a point on that same axis, representing the publicly declared stance of the individual (and thus a choice variable). The psychological cost of a type $t$ who publicly declares a stance $s$ is given by

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

$D$ can be interpreted as the cognitive dissonance or inner discomfort felt by taking a stance that does not reflect the bliss point $t$.[5] $\alpha$ captures how sensitive an individual is to small, relative to large, deviations from her bliss point, thus representing the *curvature of inner discomfort*. When $\alpha < 1$, the inner discomfort is concave, representing a meticulous or perfectionist individual attitude; when $\alpha > 1$ inner discomfort is convex, reflecting a flexibility with regard to small deviations from the bliss point.[6] An individual who takes $s$ as a stance also feels social pressure

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

Pressure arises when the stance deviates from $\bar{s}$, which can be understood as a social norm. Following the previous section, we use the labels *liberal* for $\beta > 1$ and *strict* for $\beta < 1$. $K$ represents the *weight* of social pressure relative to the psychological cost. We will assume that the only difference between individuals in society concerns their bliss points ($t$), while $\alpha$, $\beta$ and $K$ are the same for all members of society (but we will partly relax this assumption in Section 7). For conciseness we ignore the special case of $\alpha = \beta$ throughout the paper as it yields no additional insights.

[5] We use power functions for brevity and in order to facilitate the interpretation, but nearly all upcoming results can be derived using general convex and concave functional forms.

[6] Theoretically we see no particular reason why a convex or concave psychological cost function would be more or less reasonable. While in previous theoretical research a convex disutility is more common (e.g. Bernheim, 1994; Manski & Mayshar, 2003), some recent experimental research suggests concave preferences may be present in many cases too (e.g. Kendall et al., 2015; Gino et al. 2010; Gneezy et al., 2013).

The total loss (or disutility) of an individual is the sum of the inner discomfort and the social pressure.

$$L(s,t) = D(s,t) + P(s,\bar{s}) \tag{1}$$

Seeking to minimize $L(s,t)$, it is immediate that each individual will declare either her private bliss point $t$ or the social norm $\bar{s}$ (where both are corner solutions), or alternatively choose a stance strictly in between them (an inner solution). That is,

$$\forall t, s^*(t) \in \begin{cases} [\bar{s}, t], \text{ if } \bar{s} \leq t \\ [t, \bar{s}], \text{ if } t < \bar{s} \end{cases},$$

where $s^*(t)$ is the stance that minimizes the loss for type $t$. In some cases it will be useful to compare the loss incurred by the individual in the two corner solutions. This is a comparison of $D(\bar{s},t) = |t-\bar{s}|^{\alpha}$ and $P(t,\bar{s}) = K|t-\bar{s}|^{\beta}$, which boils down to comparing $|t-\bar{s}|$ and $K^{\frac{1}{\alpha-\beta}}$. Denoting $\Delta \equiv K^{\frac{1}{\alpha-\beta}}$ is then useful for the presentation of some of the results.

To compare the extent of norm conformity of different individuals in society, two different measures will be used.

**Definition 1** *The conformity of $t$ is:* $-|s^*(t) - \bar{s}|$.

This measure quantifies how close to the norm an individual's stance is. We will say that $t$ conforms more than $t'$ if $|s^*(t) - \bar{s}| \leq |s^*(t') - \bar{s}|$.

**Definition 2** *The relative concession of $t$ is:* $|t - s^*(t)| / |t - \bar{s}|$.

This measure is meant to portray the step an individual takes towards the norm when declaring a stance, compared to the step she would take if she completely conformed to the norm. We say that $t$ concedes relatively more than $t'$ if $|t - s^*(t)| / |t - \bar{s}| \geq |t' - s^*(t')| / |t' - \bar{s}|$.

The social norm $\bar{s}$ is exogenous from the point of view of an individual, but in equilibrium it will be endogenously determined by the average *stated* opinion.

$$\bar{s} = E[s^*(t)]$$

We then say that society is in equilibrium if the distribution of stances given a certain norm $\bar{s}$ has this norm as its average stance. In order to obtain the distribution of stances we also need to specify a distribution of types, which we assume is uniform, $t \sim U[t_l, t_h]$. It should be noted, however, that all the results in the paper, except for those describing the distribution of stances or the norm location, are independent of the underlying distribution of types.

Taken as a whole, the above model provides a rich description of societies (or cultures) in terms of their basic characteristics and outcomes. Each society has its own underlying characteristics consisting of the distribution of private opinions ($t$), the curvature of social pressure ($\beta$), the curvature of the individual psychological cost ($\alpha$) and the weight of pressure ($K$). In each society, we can then observe the behavior of individuals and aggregate outcomes in terms of: how conformity and concession depend on each individual's type; what the distribution of public opinions is; and what norm the society sustains.

## 4 Main patterns in strict and liberal societies

The main differences in outcomes between liberal societies ($\beta > 1$) and strict societies ($\beta < 1$) can be demonstrated using a baseline case of $\alpha = 1$.

**Proposition 1** *Suppose $\alpha = 1$. Then:*

1. ***Individual stances:*** *If society is liberal, types close to the norm speak their minds ($s^*(t) = t$) while types further away have inner solutions; if society is strict, types close to the norm fully conform ($s^*(t) = \bar{s}$) while types further away speak their minds.*

2. ***Concession:*** *If society is liberal, relative concession is weakly increasing in $|t - \bar{s}|$; if society is strict, relative concession is weakly decreasing in $|t - \bar{s}|$.*

3. ***Stance distribution:*** *If society is liberal, the distribution of stances is either uniform or bimodal; if society is strict, the distribution of stances is unimodal.*

4. ***Norm location:*** *If society is liberal, $\bar{s} = \frac{t_l + t_h}{2}$; if society is strict, any $\bar{s} \in \left\{\frac{t_l + t_h}{2}\right\} \cup [t_h - \Delta, t_l + \Delta]$ can be sustained.*

With regard to liberal societies, the first statement of the proposition expresses that moderates, who disagree only slightly with the norm, will speak their minds, while extremists, whose private views are further away from the norm, will moderate their public statements. This follows almost directly from the fact that pressure is convex, implying lenient pressure on small norm deviations. The second statement reflects this logic by saying that liberal societies achieve relatively large concessions from extremists compared to moderates. Extremists will compromise just as much as needed to reduce the most severe pressure and so will tend to bunch at a certain distance from the norm. As this happens on

both sides of the norm, the distribution will be bimodal (statement 3) and society will be polarized.[7]

Meanwhile, the meticulousness of strict societies discourages compromise – individuals will either speak their minds or completely conform (statement 1). Unlike liberal societies, strict societies will be particularly effective in getting full conformity from those who nearly agree with the norm anyway, while essentially not affecting the declarations of those who strongly disagree with it. This implies relatively large concessions by moderates (statement 2). The concentration of such people at the norm creates a unimodal distribution of stances (statement 3). This will be at the expense of the possible moderation of those who strongly disagree with the norm, who will speak their minds openly. An interpretation of this is that strict societies alienate extremists. To create cohesion, the strict society will need to use heavy pressure (large $K$) to convince extremists to fully align with the norm.

Following the fourth statement regarding norm location, the model predicts that liberal societies are bound to have norms that are representative of the actual private sentiments in society. The intuition for this is that in liberal societies everyone chooses a stance on her side of the norm and nobody fully conforms. Thus, a biased norm will imply too many declared stances on one side of the norm, preventing it from being the average stance. In contrast, in strict societies the norm may well be biased with respect to private opinions, because strict societies *can* induce the full conformity of an individual, and when this happens the individual has no effect on the norm's location (we elaborate on this in Section 6).

One interesting difference between the two kinds of societies is that the most extreme deviation from the norm will often be more extreme in strict societies than in liberal ones. This result is surprising as it can occur even when pressure is higher in strict societies. To see this, consider the following simple example. Compare two societies, one strict and one liberal, such that in both societies $t_l = -1$, $t_h = 1$ and $\bar{s} = 0$, and both have the same $K < 1$. First note that the pressure on any stance is at least as severe in the strict society as in the liberal one. Second, it is easy to show that the most extreme types, $t_{h,}$ and $t_l$, are speaking their minds ($|s| = 1$) in the strict society, while choosing a compromise stance ($|s| < 1$) in the liberal one, provided that it is liberal enough ($\beta > 1/K$).[8] In this case, the largest norm deviation will be

[7] If the distribution of types is too narrow to have extremists who compromise, the distribution of stances will be uniform.

[8] To see why $s^*(t_h) = t_h$ in the strict society, use part (1) of the proposition while noting that $L(t_h, t_h) = K < 1 = L(0, t_h)$, which means that $t_h$ prefers speaking her

observed in the strict society, even though it is harsher toward norm breakers.

The statements of Proposition 1 need to be slightly refined when considering $\alpha \neq 1$. In the next two sections we show that it is the degree of liberalism or strictness relative to the curvature of inner discomfort that drives most of the results – that is, whether $\beta$ or $\alpha$ is greater – and that the relationship between the two provides additional insights.

## 5 Liberal societies

In this section we examine the case where $\beta \geq 1$ while $\alpha$ may take any positive value. In this case there is an inner solution for every type $t$ if $\alpha > 1$ and a possibility for both inner and corner solutions if $\alpha \leq 1$. The properties of stances and conformity in liberal societies are summarized in the following proposition.

**Proposition 2** *If $\beta \geq 1$ then:*

1. *If $\beta < \alpha$, then $|s^*(t) - \bar{s}|$ is increasing and convex in $|t - \bar{s}|$, i.e., conformity is decreasing in $|t - \bar{s}|$. Moreover, the relative concession is decreasing in $|t - \bar{s}|$ and the distribution of $s^*$ is unimodal.*

2. *If $1 < \alpha < \beta$, then $|s^*(t) - \bar{s}|$ is increasing and concave in $|t - \bar{s}|$, i.e., conformity is decreasing in $|t - \bar{s}|$. Moreover, the relative concession is increasing in $|t - \bar{s}|$ and the distribution of $s^*$ is bimodal.*

3. *If $\alpha \leq 1 < \beta$ and $t_h - t_l > 2\Delta$, then $|s^*(t) - \bar{s}|$ is first increasing and then decreasing in $|t - \bar{s}|$. The relative concession is increasing in $|t - \bar{s}|$ and the distribution of $s^*$ is bimodal.*

The results are represented graphically in Figure 2 where the left panels present the function $s^*(t)$ and the right panels present the resulting distribution of stances given a uniform distribution of types.

To understand the intuition for the first part of the proposition ($\alpha > \beta$), consider a very large $\alpha$. When $\alpha$ is large, individuals feel very little dissonance if they deviate slightly from their bliss points. As illustrated in the left panel of Figure 2A, a moderate person will thus choose a stance very close to $\bar{s}$. However, an extreme type will not be willing to move as close to $\bar{s}$, as her dissonance will then be very large. Thus, in relative terms, extremists tend to concede less than moderates. The resulting

mind. To see why $s^*(t_h) < t_h$ in the liberal society, note that the derivative of the loss function for type $t_h$ at $s = t_h = 1$ is given by $L'(t_h, t_h) = -1 + \beta K (t_h - \bar{s})^{\beta - 1} = \beta K - 1 > 0$, which means that this type has a profitable deviation from $s = t_h = 1$ to $s < 1$.

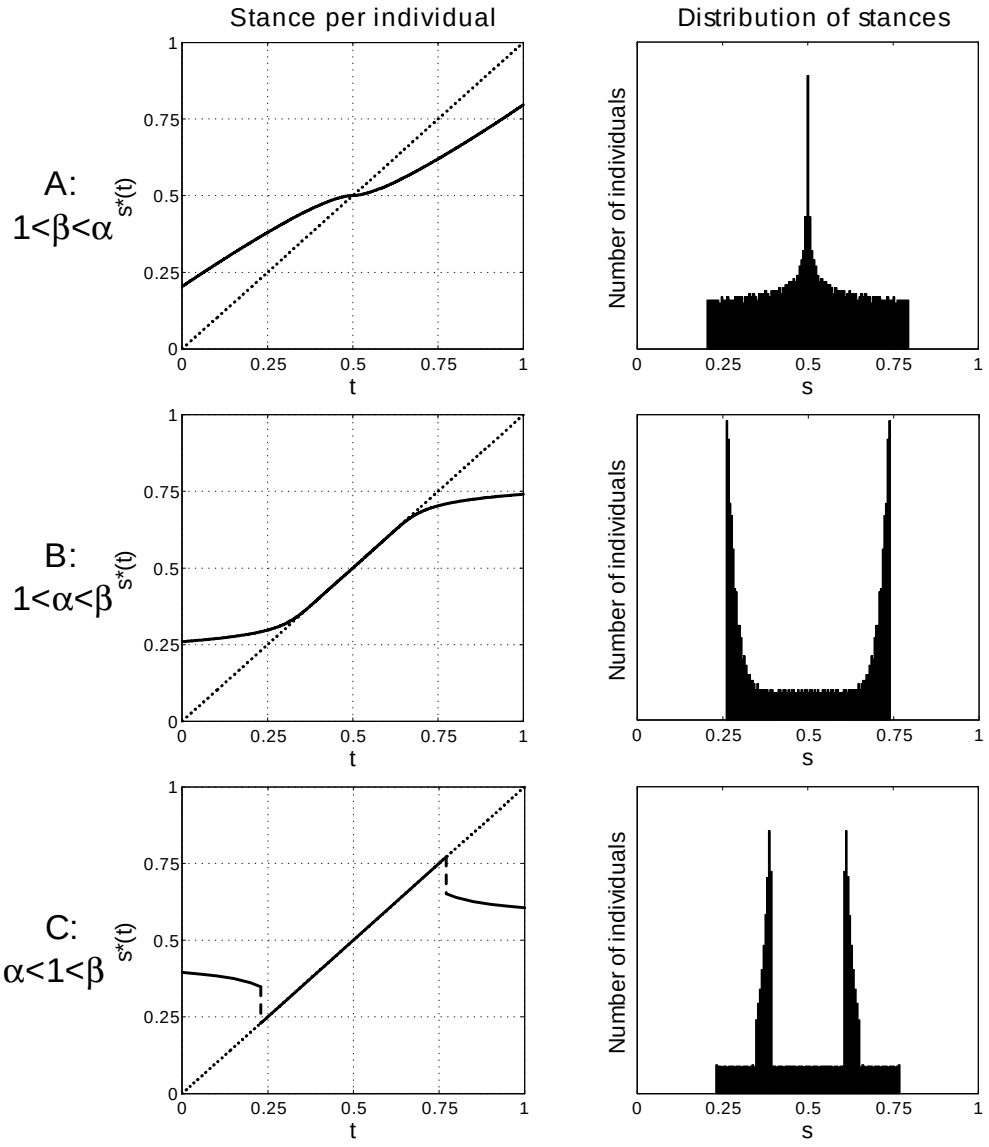


Figure 2: The left panels depict $s^*(t)$ (full line) and $s = t$ (dashed line). The right panels depict the distribution of stances. In all graphs $t \sim U(0,1)$ and $\bar{s} = .5$. Panel A: $K = 0.5$, $\alpha = 2$, $\beta = 1.2$. Panel B: $K = 2$, $\alpha = 1.1$, $\beta = 2$. Panel C: $K = 2$, $\alpha = 0.85$, $\beta = 1.5$.

distribution of stances (right panel) will therefore be a concentration of statements around the norm, created by the moderate types.

To understand the intuition for the second and third parts of the proposition ($\beta > \alpha$), consider now a very large $\beta$. A large $\beta$ implies that individuals feel very little pressure when they deviate a little from $\bar{s}$, but the pressure rises steeply when the deviation from $\bar{s}$ is large. Consequently, moderates will move only slightly from their bliss points, if at all (left panels of Figures 2B and 2C). Meanwhile, extreme types will take large steps from their bliss points, due to the high social pressure on large deviations from the norm. The result will be a concentration of extreme types at a certain distance from the norm on each side of it and society will look polarized, as is illustrated in the right panels of Figures 2B and 2C. The baseline case of $\alpha = 1$ is a special case of this (see Proposition 1).

There is, however, an important twist to liberal societies when inner discomfort is concave ($\alpha < 1 < \beta$). When individuals are sensitive to small deviations from their bliss points, they will tend to either speak their minds or, once they deviate from their bliss points, state almost anything that lowers social pressure. Since in liberal societies pressure is convex, moderate individuals will be under low pressure and hence not make any concessions. Meanwhile, extremists would be under high pressure if they spoke their minds. Therefore, they will be forced to concede, and as $\alpha < 1$, these concessions will be quite extensive, implying that extremists will conform even more than some moderates – a pattern that may be called *inversion of opinions*.[9] As a result, conformity will be non-monotonic in the distance from the norm, as illustrated in the left panel of Figure 2C.[10]

In the baseline case of Section 4 we saw that liberal societies are mainly effective in inducing conformity among extremists and that this leads to a bimodal distribution of stances. The analysis in this section makes an important refinement to these results. If the degree of liberalism is high (i.e., $\beta > \alpha$), society is indeed mainly directed at inducing conformity by extremists, which leads to bimodality. On the other hand, a low degree of liberalism (i.e., $\beta < \alpha$) induces conformity by moderates, leading to a unimodal concentration.

[9] In fact, we get inversion at two levels. Firstly, between extremists and moderates, the extremists conform more than some moderates. Secondly, within the group of extremists, the most extreme conform more than the less extreme.

[10] The third part of the proposition considers only the case where the distribution of types is sufficiently broad, $t_h - t_l > 2\Delta$. If the distribution of types is too narrow to have extremists who compromise, the distribution of stances will be uniform.

## 6 Strict societies

In this section we examine the case where $\beta \leq 1$ while $\alpha$ can take any positive value. In this case, if $\alpha \leq 1$, any inner solution to the individual's minimization problem is a maximum, implying that individuals will either fully conform or speak their minds. This is intuitive, as an individual with concave discomfort in a strict society, who takes a stance in-between $t$ and $\bar{s}$, would feel both great inner discomfort and heavy pressure. When $\alpha > 1$, there is a possibility for both inner and corner solutions. The properties of stances and conformity in a strict society are summarized in the following proposition.[11]

**Proposition 3** *If $\beta \leq 1$ and $t_h - t_l > 2\Delta$, then:*

1. *If $\beta < \alpha \leq 1$, then types with $|t - \bar{s}| < \Delta$ fully conform while types with $|t - \bar{s}| > \Delta$ speak their minds. Conformity and relative concession are weakly decreasing in $|t - \bar{s}|$. The distribution of $s^*$ has a peak at $\bar{s}$ and uniform tails at the extreme ends of the range.*

2. *If $\alpha < \beta$, then types with $|t - \bar{s}| < \Delta$ speak their minds while types with $|t - \bar{s}| > \Delta$ fully conform. Conformity is first decreasing in $|t - \bar{s}|$ and then sharply increases. Relative concession is weakly increasing in $|t - \bar{s}|$. The distribution of $s^*$ has a peak at $\bar{s}$ and a uniform section attached to it.*

3. *If $\beta < 1 < \alpha$, then there exists a cutoff distance $\delta < \Delta$ such that types closer than $\delta$ to the norm fully conform, while types further than $\delta$ from the norm have inner solutions. Conformity and relative concession are weakly decreasing in $|t - \bar{s}|$. The distribution of $s^*$ is discontinuously trimodal with a central peak at $\bar{s}$ and a detached part on each side.*

In part 1 of the proposition, society displays a relatively high degree of strictness ($\beta < \alpha \leq 1$). It is illustrated in Figure 3A. Here moderates choose to fully conform, while extremists simply cope with the full social pressure and express their bliss points. This happens because in strict societies one has to move all the way to the norm to alleviate pressure to a substantial degree. Thus, when $\beta < \alpha$, extreme types find it relatively more painful to move to the norm compared to speaking their minds. Overall, this means that very strict societies alienate people with

[11] The proposition considers only the case where the distribution of types is sufficiently broad. If the distribution of types is narrow, then, when $\beta < \alpha$, this would lead to full conformity by all individuals and when $\alpha < \beta$, it will lead to all individuals speaking their minds.

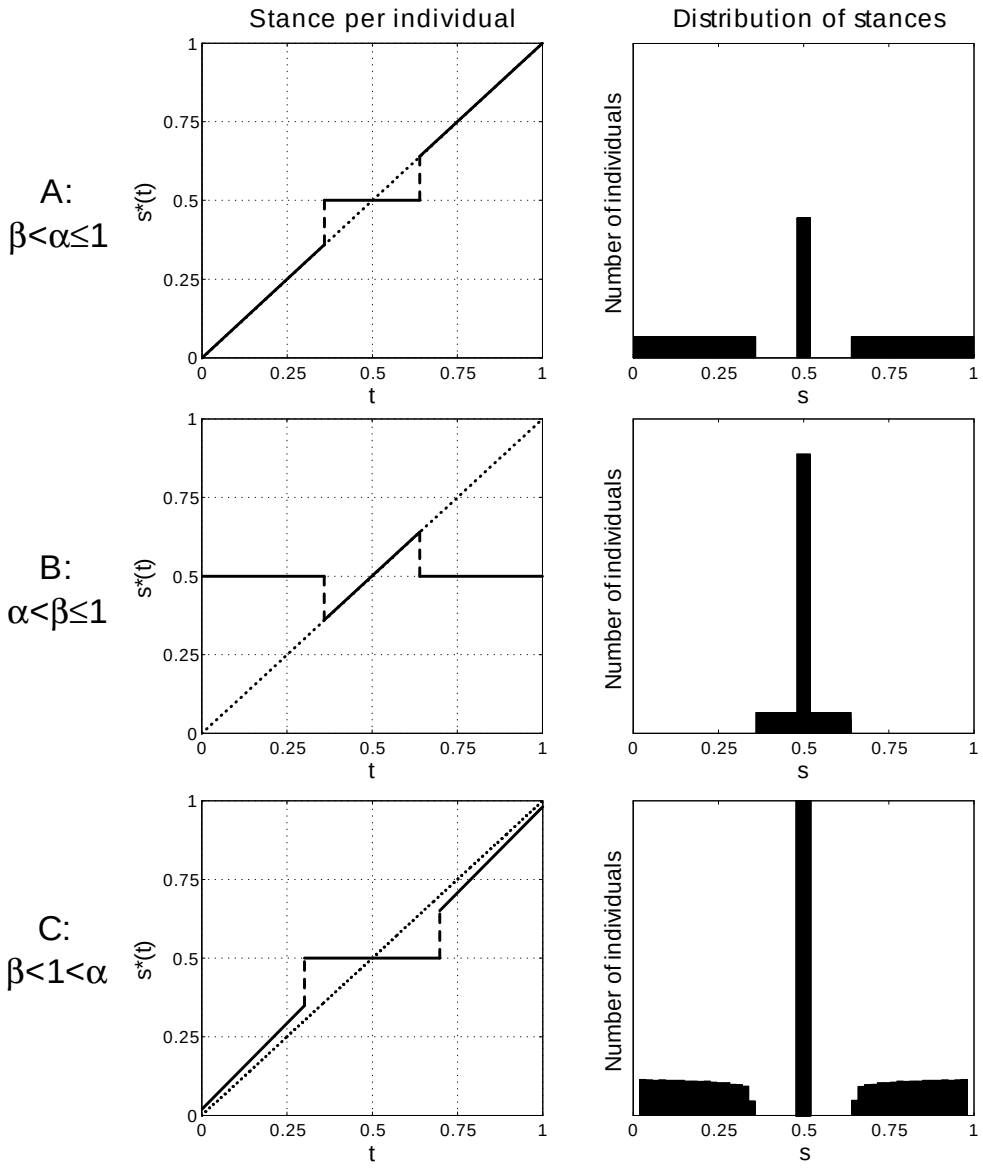


Figure 3: The left panels depict $s^*(t)$ (full line) and $s = t$ (dashed line). The right panels depict the distribution of stances. In all graphs $t \sim U(0,1)$ and $\bar{s} = .5$. Panel A: $K = 0.5$, $\alpha = 0.85$, $\beta = 0.5$. Panel B: $K = 2$, $\alpha = 0.5$, $\beta = 0.85$. Panel C: $K = 0.5$, $\alpha = 1.25$, $\beta = 0.8$.

opinions far from the norm, but compel those with opinions close to the norm to fully conform.

When society is strict to a lesser degree, so that $\beta > \alpha$ (part 2 of Proposition 3), extreme types fully conform to the norm while moderates speak their minds. The intuition is that moderates are unwilling to conform as the psychological cost to them, of deviating from their bliss point, is very concave. For extremists, however, not conforming will imply a great deal of social pressure. All in all, this kind of society will be good at attracting extremists to the norm while "allowing" freedom of expression of those (sufficiently) close to it.[12] The observable outcome of this case is a distribution that looks like a standard concentration of individuals at and around the norm (right panel of Figure 3B). But there is an important twist. Here we get the pattern of inversion of opinions, where those who despise the norm the most are the (only) ones conforming.

By comparing part 1 and part 2 of the proposition, we see that both kinds of strict society with concave discomfort have one thing in common – they foster an all-or-nothing mentality, making each person either conform fully or not at all. But the degree of strictness leads to an important refinement as it yields further predictions that are completely opposite depending on whether $\beta > \alpha$ or $\beta < \alpha$. Firstly, less strict societies ($\beta > \alpha$) are predicted to induce extremists to conform, while stricter societies ($\beta < \alpha$) are predicted to induce moderates to conform. Secondly, less strict societies are predicted to have a unimodal concentration around the norm (right panel of Figure 3B), while stricter societies are predicted to have a peak at the norm with detachment at the extremes (right panel of Figure 3A).

In the third part of the proposition, when $\alpha > 1$, only large deviations from the bliss point create inner discomfort. We then get a combination of corner and inner solutions, where individuals with opinions far enough from the norm choose an inner solution, while moderates completely conform to the norm. The right panel of Figure 3C illustrates the resulting distribution of stances. The distribution has a peak at $\bar{s}$ and a detached part towards each of the extreme ends. The intuition for this is that, as society is strict, small deviations from the norm draw relatively heavy pressure. When this is the case and individuals perceive small deviations from their bliss points as almost painless, moderates do best by completely conforming to the social norm. In comparison, because of the convexity of $D$, extremists would feel too much discomfort

[12] In fact, for any finite $K$, no matter how large, there will always be a group of types close to the norm who speak their minds. Hence, full conformity by all cannot be attained here.

if they were to fully conform. However, they do not mind making small concessions, and hence choose a compromise solution. Just like the case in which individuals have concave discomfort and society is very strict ($\beta < \alpha < 1$), here too there is alienation, although extremists do make some concessions. The general lesson from these two cases is that extreme strictness alienates those who strongly disagree with the norm. One interpretation is that these individuals would prefer to live outside the community, as in the Jewish Ultraorthodox community, which practices excommunication.

We turn now to analyze which norms a strict society can sustain. In Section 4 (Proposition 1, statement 4) we stated that liberal societies can sustain only a central norm, while strict societies can also sustain a norm that is biased with respect to private opinions. That is, the norm in strict societies may be unrepresentative of the underlying preferences in society. The following proposition outlines the conditions under which this might happen.

**Proposition 4** *If $\beta \leq 1$, there exists an equilibrium where $\bar{s} = \frac{t_l + t_h}{2}$. Furthermore:*

1. *If $\beta < \alpha$, there exist equilibria with $\bar{s} \neq \frac{t_l + t_h}{2}$ if and only if $t_h - t_l < 2\delta$, where $0 < \delta \leq \Delta$ .*

2. *If $\alpha < \beta$, there exist equilibria with $\bar{s} \neq \frac{t_l + t_h}{2}$ if and only if $t_h - t_l > 2\Delta$.*

Part 1 of the proposition highlights that when $\alpha$ is larger than $\beta$, a biased norm requires a narrow range of types. As we previously saw in Proposition 3, when $\beta < \alpha$ moderates fully conform. The requirement for a narrow range of types in part 1 of Proposition 4 is there to ensure that society consists only of such moderates, i.e., that all types fully conform. Otherwise a biased norm would lead to a greater mass of opposition at one of the extreme ends than the other, implying the norm is not the average stance. The need for full conformity implies that, in order to uphold a biased norm, very strict societies require cohesion of statements. This can occur either if private preferences are themselves cohesive (i.e., there is a narrow range of bliss points) or if severe social pressure is employed, which induces artificial cohesion of statements.

The second part of Proposition 4 deals with the case in which moderates speak their minds while extremists fully conform (see Figure 3B). By fully conforming, extremists give up their say in determining the norm's location, thus enabling it to be unfavorable to them. This implies that, unlike the case of $\beta < \alpha$, here it is not necessary to have cohesive private

opinions in order to sustain a biased norm. In fact, a broader range of types enables a broader range of norms. This is so because the norm has to balance only the non-conforming statements expressed by those close to it – individuals within distance $\Delta$ from the norm. A broader range of types – i.e., less cohesive private opinions – enables the norm to be located further away from the center of the type distribution without unbalancing these non-conforming statements.

# 7 Testable implications and further results

This section highlights some of the model predictions and presents several further results that can be empirically tested. Each subsection contains an empirical prediction, the intuition behind it and a description of methods and sources of data that can be used for testing it. In subsection 7.1 we also present an illustrative test of the prediction in question. The predictions presented in subsections 7.1 to 7.3 are independent of the distribution of types and hence do not require knowing it. They are also independent of $\alpha$ and can therefore be tested even when $\alpha$ is unknown or heterogeneous in society. Section 7.4 presents a prediction that *is* dependent on the type distribution and on $\alpha$ and provides applications where it may be tested.

## 7.1 The effect of $\beta$ on full conformity and maximum norm deviation

Throughout the paper, we have highlighted that strict societies tend to facilitate the choice of corner solutions by the individual, while liberal societies tend to facilitate inner solutions. The following proposition can be used to empirically test this claim. For that purpose, we denote by $s_h$ the stance that constitutes the maximal feasible deviation from the norm (to the right).

**Proposition 5** *Suppose $s_h - \bar{s} = 1$. Then the proportion of individuals choosing $s \in \{\bar{s}, s_h\}$ is weakly decreasing in $\beta$.*

By normalizing $s_h - \bar{s}$ to 1 we are essentially fixing the pressure on the maximal deviation to $K$. Hence, the proposition considers the pure effect of a change in the curvature of pressure ($\beta$) on the statements made in equilibrium. It essentially says that, *ceteris paribus*, a stricter society will show larger concentrations of stances at the edges of the distribution (at $\bar{s}$ and at $s_h$). The intuition for this is straightforward. When decreasing $\beta$, the social pressure at $\bar{s}$ and $s_h$ is unchanged, while it increases at all intermediate stances. This means that full conformity and maximal deviation become relatively more attractive, thus increasing the share

of individuals taking such stances. The result is independent of the $t$-distribution, $\alpha$ and $K$. Naturally, societies may differ with respect to these parameters. However, as long as these parameters are either uncorrelated with $\beta$ or can be controlled for in a regression, the proposition gives a testable prediction.

As an example, consider religious practice in relation to a religious norm. The sixth wave of the World Value Survey (WVS) measures public support for statements such as: "the only acceptable religion is my religion" (V154) and "people who belong to different religions are probably just as moral as those who belong to mine" (V156).[13] The support for these statements can be thought of as measuring $\beta$. For instance, agreement with "the only acceptable religion is my religion" suggests an intolerance towards small deviations from one's own religion or extent of religiosity – a small $\beta$. According to Proposition 5, the smaller $\beta$ is in a society, the larger should be the total share of individuals either fully adhering to the norm or maximally disobeying it. Thus, one can collect data on religious norms and religious practices across countries and test the prediction using the proxy for $\beta$ obtained from the WVS.

As a rough illustration of how the test can be performed, consider the answers to the WVS question (V146) "How often do you pray?". In Muslim societies, the norm can be plausibly assumed to be five times a day, as this is the commandment stated in the Koran. In the WVS, this would be reported as the maximal frequency of praying ("several times a day"). The largest possible deviation from the norm is not to pray at all, which corresponds to answering "never, practically never" in the survey. Thus, testing our prediction using this question is straightforward. If the prediction is correct, then $\beta$ should be *negatively* correlated with the share of individuals reporting either of these two extremes. Likewise, $\beta$ should be *positively* correlated with the share of individuals reporting intermediate frequencies of praying.

The WVS includes 16 countries in which the major religion is Islam and for which the necessary data is available (see the list and further details in the appendix). We ran a simple regression with no controls to examine the fraction of people who report either the maximum or the minimum frequency of prayers as a function of our proposed measure of $\beta$.[14] The left panel of Figure 4 displays the data and regression

[13] The WVS has been used extensively in the economics literature to measure cultural traits, values and norms. See Knack and Keefer (1997) for an early application. Like in the previous literature, we treat answers in the WVS as being truthful.

[14] More precisely, our measure of $\beta$ is the share of people who, with respect to the statement "The only acceptable religion is mine", answer "strongly disagree", minus

graph. As predicted, the slope of the regression line is negative (it is also statistically significant). This result is further corroborated by the positive (and significant) correlations between our proxy for $\beta$ and each intermediate extent of prayer. One such example is depicted in the right panel of Figure 4. It shows that the share of individuals praying "once a day" to "several times a week" is positively correlated with the $\beta$ proxy.[15] The results suggest that stricter societies are not necessarily making people behave more religiously in general, but are specifically effective in fostering corner solutions.[16]

As a simple robustness check, we used the answers to two additional WVS questions as controls. We chose these specific controls in order to address concerns that our results might simply be driven by strict societies being more religious and harsher toward religious deviations. To measure the degree of religiosity, we used the share of individuals who declared they were religious (V147). This can be thought of as a proxy for the location of the type distribution in religious space. The extent of agreement with the statement "An essential characteristic of democracy is: Religious authorities ultimately interpret the laws" (V132) can be thought of as a proxy for $K$.[17] All of the previous results were replicated. The regression tables are reported in the appendix. Given that the tests are based on cross country data using a small number of countries, these results should of course be interpreted with caution. Furthermore, the results merely show correlations and do not establish causality. Preferably, a more extensive test would use the proxy for $\beta$ from the WVS in order to test the prediction in a natural or a field experiment. Alternatively, it would be useful to control for religiosity and other factors on

the share answering "strongly agree".

[15] The WVS question on frequency of praying contains 10 possible answers, two of which are "no answer" and "don't know". As for the other eight possible answers, we took the intermediate six, ranging from "once a day" to "less often than once a year", and divided them into three pairs of adjacent answers (corresponding to answer codes 2-3, 4-5, 6-7). For each such pair we found a positive and significant slope for the fraction of people reporting an answer within the pair as a function of our measure of $\beta$. Figure 4 (right) presents the regression graph for the first pair. The regression tables for all pairs can be found in the appendix.

[16] The correlation between our measure of $\beta$ and the share of people who report the maximum frequency of praying ("several times a day") is significantly negative, while the correlation between our measure of $\beta$ and the share of people who report the minimum frequency of praying ("never, practically never") is insignificant and very close to zero. The former is thus in line with the model, while the latter neither corroborates nor refutes it. The latter does, however, further corroborate that strictness, as we measure it, does not simply imply more praying in general, as otherwise the share of those never praying should have been lower in stricter societies.

[17] The scale goes from 1 ("Not an essential characteristic of democracy") to 10 ("An essential characteristic of democracy"). We use the mean.

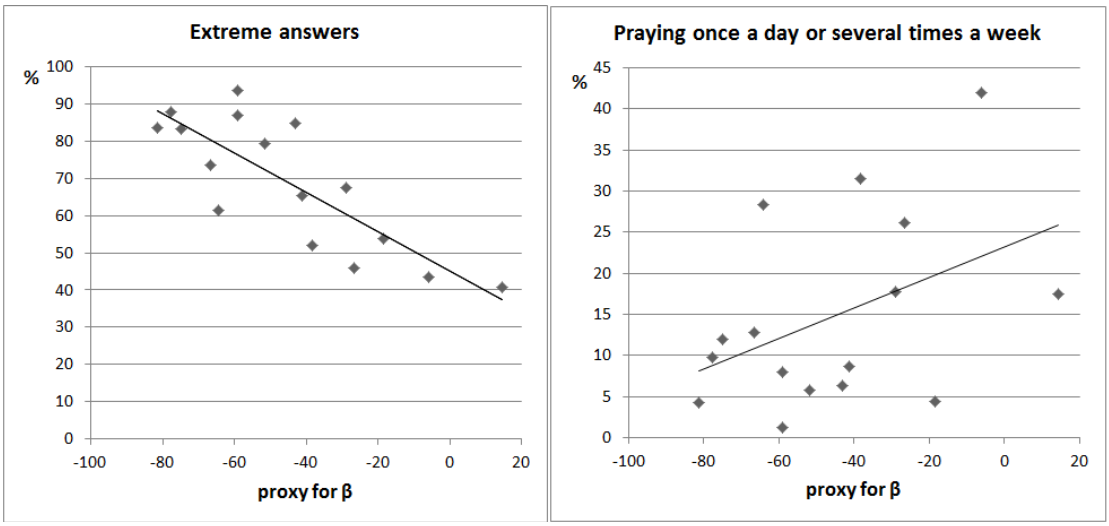


Figure 4: Data and estimated linear correlation between the proxy for $\beta$ from answers to WVS V154 (share of people answering "strongly disagree" minus share answering "strongly agree") and reported extent of praying (WVS V146). The left panel displays the share of people praying "several times a day" or "never, practically never". The right panel displays the share of people praying "once a day" or "several times per week".

an individual level. The purpose here has mainly been to illustrate how the model could be tested. However, at least as a first pass, we do find the results to be a sign of the potency of the model to generate valid predictions.

## 7.2 The effect of $K$ on full conformity

An implication of the model that follows directly from Propositions 2 and 3 and their proofs is the following.

**Proposition 6** *An increase in $K$ (weakly) increases the number of individuals stating $\bar{s}$ if and only if $\beta \leq 1$.*

The proposition predicts that increasing the weight of pressure should produce a higher number of full conformers only in strict societies. In liberal societies, such an increase may shift people in the direction of the norm, but will not induce full conformity. Here again, the tendency of strict societies to facilitate the choice of corner solutions and the tendency of liberal societies to facilitate the choice of inner solutions jointly determine the result.

To see how this prediction can be tested, consider the dictator game. Krupka and Weber (2013) measured the social appropriateness of divisions in the dictator game and found not only that an equal split is the most socially appropriate division (thus constituting a norm), but also that the social pressure on deviations from an equal split is concave. In light of Proposition 6, this suggests that, following an increase in the prominence of social pressure in the dictator game, we should observe an increase in the number of individuals choosing an equal split. In order to test this prediction, one can manipulate the effect of $K$ in the dictator game by conducting the same experiment, once under anonymity and once under full transparency, and then observe how divisions change between the two settings. Alternatively, one can vary the perceived probability that subjects are observed by others.

An alternative application would be to test the prediction on sensitive issues such as sexual preference or political opinion. To do this, the first step would be to elicit the perceived social appropriateness (i.e., the norm and the curvature of social pressure) of certain preferences and opinions using the technique of Krupka and Weber (2013).[18] The next step would be to ask (a new set of) respondents to state their own preferences ($s$) while varying $K$ across treatments. In order to vary $K$, one can use different degrees of anonymity in the survey. One simple way to do this would be to use the *randomized response technique*. Here, respondents would roll a die that decides whether they should answer truthfully or answer randomly with the help of a new roll of the die.[19] By varying the probability of having to give a truthful answer, the survey maker can vary $K$: a higher probability of having to answer truthfully implies a higher $K$. The proposition then predicts that, as we increase $K$, we should see a stronger effect of clustering at $\bar{s}$ in groups and issues for which the initial step showed that $\beta \leq 1$, compared to groups and issues for which $\beta > 1$.

[18] The problem of identifying norms and pressure in surveys has been that responses may be confounded by what the responders personally think is right and wrong. To get a more direct measurement of collectively perceived appropriateness, Krupka and Weber (2013) ask individuals anonymously for the social appropriateness of different behaviors and reward them for matching the responses of others.

[19] The randomized response technique (Warner, 1965) was devised in order to elicit private information or attitudes on sensitive issues. In the original usage of the technique, subjects privately flip a coin before answering a binary question to which a positive answer is regarded sensitive, and are instructed to answer "yes" if the coin comes up tails and truthfully if it comes up heads. This method does not enable the experimenter to get data on the individual level, but aggregate distributions can be easily elicited. It is straightforward to extend this technique in order to enable non-binary responses by, for instance, using a die instead of a coin as we suggest here.

## 7.3 The effect of $\beta$ on small vs. large norm deviations

The two previous subsections dealt with the effect of pressure on the tendency to choose full conformity and maximum deviation from the norm. This subsection deals with the effect of pressure on relatively small vs. relatively large deviations from the norm. As will be shown, these effects can be important from the point of view of a legislator.

**Proposition 7** *Let $S_{low}$ denote the set of all observed stances in $]\bar{s}, \bar{s}+1]$ and let $S_{high}$ denote the set of all observed stances above $\bar{s}+1$. Then, holding $\bar{s}$ constant, $|S_{low}| - |S_{high}|$ weakly increases in $\beta$.*

The proposition examines the distribution of stances to the right of the norm and compares the proportion of stances "close" to it ($s \in S_{low}$) to the proportion "far" from it ($s \in S_{high}$). An increase in $\beta$ implies that the pressure decreases in $S_{low}$ while it increases in $S_{high}$, leading individuals to have a higher tendency to choose stances in $S_{low}$ and a lower tendency to choose stances in $S_{high}$ (so that $|S_{low}| - |S_{high}|$ increases). The cutoff between the two regions is at distance 1 from the norm, so that $s = \bar{s} + 1$ is a "flexpoint", around which the pressure function rotates as $\beta$ increases.[20]

This result has important implications for legal sanctions. Enforcing laws under a given budget constraint often implies a trade-off between catching small and large offenders. For example, if the police try to stop every driver they see exceeding the speed limit even slightly, this will tend to lower the probability of catching the very fast drivers. In practical terms, such a policy resembles a lowering of $\beta$. The testable prediction of the proposition is that such a change in enforcement will lead to a decrease in small deviations from the speed limit but an increase in large deviations.

## 7.4 Relative concession across individuals

In Definition 2 we introduced *relative concession* as a measure for the step an individual takes towards the norm when declaring a stance, compared to the step she would take if she completely conformed. Part 2 of Proposition 1 then states that, when $\alpha = 1$, relative concession increases in $|t - \bar{s}|$ when pressure is convex and decreases in $|t - \bar{s}|$ when

[20] Note that unlike Proposition 5, where we set $s_h - \bar{s} = 1$ in order to fix the maximal pressure, Proposition 7 has interesting implications when $s_h - \bar{s} > 1$, so that changing $\beta$ affects pressure on small and large deviations differently. The flexpoint itself can be generalized to any location. This requires a small manipulation of the pressure function to $P = K\left(A\left(s - \bar{s}\right)\right)^{\beta}$, with $s = \bar{s} + 1/A$ being the flexpoint.

pressure is concave. As the case of $\alpha = 1$ may reasonably apply to settings in which the cost of deviation is monetary, this result yields a testable prediction in the realm of tax evasion.

The true income of all individuals is often unobservable. But for those who are audited, the authorities usually conduct a thorough investigation of the actual tax due. In terms of our model, the tax due can be thought of as the value of $t - \bar{s}$ (where $\bar{s}$ can be set to zero, reflecting that all taxes should be paid). The tax an individual evades is the equivalent of $s - \bar{s}$ in the model. The relative concession thus captures the share of tax due that was actually paid by an individual ($\frac{t-s}{t-\bar{s}}$). Hence, the prediction is that a convex sanctioning of tax evasion will lead heavy debtors to pay a higher share of their taxes than small debtors. Concave sanctioning, on the other hand, will lead small debtors to pay most of their taxes, while achieving low compliance among heavy debtors.[21]

The next corollary is a generalization of the result about relative concession to any value of $\alpha$.

**Corollary 8** *The relative concession is increasing in* $|t - \bar{s}|$ *if and only if* $\beta > \alpha$.

This generalization follows directly from Propositions 2 and 3.[22] If one has access to both the private ($t$) and the public ($s$) opinions of individuals in a certain setting, this corollary can be tested as follows. First, standard methods of structural estimation enable one to jointly estimate $\alpha$, $\beta$ and $K$ using our theoretical mapping from $t$ to $s^*(t; \alpha, \beta, K)$. This allows inferring which of $\alpha$ and $\beta$ is the greatest. Then, Corollary 8 can be tested by examining changes in the relative concession of an individual as a function of changes in $\bar{s}$ (i.e., changes in the peer composition around her).

To see how a test can be implemented, consider for instance students' achievements in school, where the class composition can be expected to

[21] Note that, in practical enforcement of tax rules, there is a difference between the auditing rule (which is based on the observed taxes paid) and the fine imposed (which is applied following an audit and is based on how much a person evaded). The expected fine for evading a certain amount of tax is a non-trivial combination of the two parts. Most standard theoretical models abstract from this distinction (see Slemrod, 2001, or Slemrod and Yithaki, 2002, for an overview) and assume individuals have a perception of the expected fine for a given level of tax evasion. One may estimate this perception through a survey.

[22] It can be generalized even further. With general functional forms, the condition for decreasing relative concession is $\gamma_P(x) \equiv -\frac{xP''(x)}{P'(x)}|_{|s^*(t)-\bar{s}|} > \gamma_D(x) \equiv -\frac{xD''(x)}{D'(x)}|_{|t-s^*(t)|}$. Here $\gamma_F(x)$ is the Arrow-Pratt measure of relative risk aversion of the function $F(x)$.

determine whether a high or a low grade is considered normative. In the recent literature on peer effects in schools (see, e.g., Leuven et al, 2010; Leuven & Rønning, forthcoming), pre-treatment test scores are used as a proxy for ability (i.e., $t$), while post-treatment test scores are used as a proxy for behavior (i.e., $s$). The empirically observed and the theoretically predicted mapping from abilities to scores can then be used to structurally estimate $\alpha$, $\beta$ and $K$, where the average post-treatment test score in a class forms the norm. Next, in a field experiment in the spirit of Booij et al. (2015), students can be repeatedly assigned to tutorial groups of different compositions (i.e., having different norms). One can then measure how close a student's score in a certain class is to the average score in that same class. The test of the model would consist in seeing whether this measure is changing, across classes and individuals, in a way that is consistent with our predictions. Suppose, for example, that according to the structural estimation $\beta > \alpha$, and consider the distance between the pre-treatment score ($t$) of a low ability student and the average score in her group ($\bar{s}$). The corollary predicts that this student should close a larger share of that distance when placed with high ability peers compared to when placed with medium ability peers.[23]

## 8 Conclusion

This paper has presented a simple theory of how social pressure affects the distribution of stated opinions and visible actions across societies. The main message is that the curvature of social pressure, and how it relates to the curvature of individuals' inner discomfort when deviating from their bliss points, is important when considering individual conformity across societies. Drawing on observations of sanctioning in different societies and cultures, both experimental and informal, we applied labels to the curvature of social pressure in order to connect the results of the model to outcomes across societies: strict societies are those emphasizing complete adherence to their code of conduct, hence utilizing concave social pressure; liberal societies are those allowing freedom of expression as long as it is not too extreme, hence utilizing convex social pressure.

In liberal societies, the convexity of the social pressure naturally in-

[23] There are other relevant settings in which $t$ and $s$ can be observed and a similar approach may be used. For instance, in a study of obesity, Carrell et al. (2011) distinguish between pre-treatment fitness ($t$) and post-treatment fitness ($s$), after individuals have been subjected to peer effects. Similarly, in law and economics it is common to use exogenous ideological scores as a proxy for judges' private political preferences (see, for instance, Epstein et al. 2007), which can then be compared to their rulings in court ($s$) when they are under pressure to make a unanimous decision (see Epstein et al. 2011 for evidence on collegial pressure in courts).

duces individuals to compromise between fully conforming and stating their private opinions in public. However, depending on the degree of liberalism – the convexity – the distribution of declared stances will be either a bimodal polarization or a unimodal concentration. In strict societies, the concavity of the social pressure discourages compromise. That is, it will tend to induce individuals to either completely conform or completely speak their minds. Depending on the degree of strictness, the conformists in society will be those whose private opinions are either quite close to the norm anyway or, rather surprisingly, quite far from it. The latter case displays inversion of opinions at the aggregate level of society, as those who dislike the norm the most adhere to it more than others.

Another prediction of the model is that liberal societies are bound to have social norms that are representative of the average private opinion in society – biased norms cannot be sustained in equilibrium. This may be linked to the informal observation that a liberal atmosphere often correlates with democracy. The model further predicts an association between strict societies and biased norms – only in strict societies is it possible to sustain a biased norm.

## References


[1] Akerlof, G. A. (1980). "A theory of social custom, of which unemployment may be one consequence". *The Quarterly Journal of Economics*, Vol. 94, No. 4, pp.749-775.

[2] Bénabou, R., & Tirole, J. (2006), "Incentives and prosocial behavior," *American Economic Review*, Vol. 96, No. 5, pp. 1652-1678.

[3] Bénabou, R., & Tirole, J. (2011), "Laws and norms". National Bureau of Economic Research (No. w17579).

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

[5] Brock, W.A., Durlauf, S.N., (2001), "Discrete Choice with Social Interactions", *Review of Economic Studies* Vol. 68, Iss. 2, pp. 235–260.

[6] Booij, A., Leuven, E., & Oosterbeek, H. (2015). "Ability peer effects in university: Evidence from a randomized experiment". IZA Discussion Paper No. 8769.

[7] Carrell, S. E., Hoekstra, M., & West, J. E. (2011). "Is poor fitness contagious?: Evidence from randomly assigned friends". *Journal of Public Economics*, Vol. 95, No. 7, pp. 657-663.

[8] Clark, A. E., & Oswald, A. J. (1998). "Comparison-concave utility and following behaviour in social and economic settings." *Journal of Public Economics*, Vol 70, Iss. 1. , pp 133-155.

[9] Eguia, J.X. (2013). "On the Spatial Representation of Decision Profiles." *Economic Theory*, Vol. 52, Iss. 1, pp 103-128.

[10] Epstein, L., Martin, A. D., Segal, J. A., & Westerland, C. (2007). "The judicial common space". *Journal of Law, Economics, and Organization*, Vol. 23, No. 2, pp. 303-325.

[11] Epstein, L., W. M. Landes, and R. A. Posner (2011): "Why (and When) Judges Dissent: A Theoretical and Empirical Analysis," *Journal of Legal Analysis*,Vol. 3, Iss. 1, pp. 101-137.

[12] Gino, F., Norton, M. I., & Ariely, D. (2010). "The Counterfeit Self The Deceptive Costs of Faking It." *Psychological Science*, Vol. 21, No. 5, pp. 712-720.

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

[14] Herrmann, B., Thöni, C., & Gächter, S. (2008). "Antisocial Punishment across Societies." *Science*, Vol. 319, No. 5868, pp. 1362-1367.

[15] Kendall, C., Nannicini, T., & Trebbi, F. (2015). "How do voters respond to information? Evidence from a randomized campaign", *American Economic Review*, Vol. 105, No. 1, pp. 322-353.

[16] Knack, S., & Keefer, P. (1997). "Does social capital have an economic payoff? A cross-country investigation". *The Quarterly journal of economics,* Vol. 112, No. 4, pp. 1251-1288.

[17] Krupka, E. L., & Weber, R. A. (2013). "Identifying social norms using coordination games: Why does dictator game sharing vary?". *Journal of the European Economic Association*, Vol. 11, No. 3, pp. 495-524.

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

[19] Kuran, T., & Sandholm, W. H. (2008). "Cultural integration and its discontents". *The Review of Economic Studies*, Vol. 75, No. 1, pp. 201-228.

[20] Leuven, E., Oosterbeek, H., & Klaauw, B. (2010). "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment" *Journal of the European Economic Association*, Vol. 8, No. 6, pp. 1243-1265.

[21] Leuven, E., & Rønning, M. (forthcoming). "Classroom grade composition and pupil achievement". *Economic Journal*

[22] 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.

[23] Lopez-Pintado, D., & Watts, D. J. (2008). "Social influence, binary

decisions and collective dynamics". *Rationality and Society*, Vol. 20, No. 4, pp. 399-443.
[24] 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.
[25] Michaeli, M. & Spiro, D., (2014), "From Peer Pressure to Biased Norms: Formation and collapse", Dept. of Economics, University of Oslo WP series Memo 15/2014.
[26] Slemrod, J. (2001). ´´A general model of the behavioral response to taxation". *International Tax and Public Finance*, Vol. 8, No. 2, pp. 119-128.
[27] Slemrod, J., & Yitzhaki, S. (2002). "Tax avoidance, evasion, and administration". *Handbook of public economics*, Vol. 3, pp. 1423-1470.
[28] Warner, S. L. (1965). "Randomized response: A survey technique for eliminating evasive answer bias". *Journal of the American Statistical Association*, Vol. 60, No. 309, pp. 63-69.

# A Appendix – Proofs and derivations

## A.1 Some useful results

### A.1.1 Conformity and relative concession

Minimizing (1) and by way of the implicit function theorem, we get the following derivatives of $s^*(t)$:

$$\frac{ds^*}{dt} = \frac{D''\left(t-s^*\right)}{P''\left(s^*\right)+D''\left(t-s^*\right)} \tag{2}$$

$$\frac{d^2s^*}{dt^2} = \frac{\left[D'''\left(t-s^*\right)\left(P''\left(s^*\right)\right)^2 - P'''\left(s^*\right)\left(D''\left(t-s^*\right)\right)^2\right]}{\left(P''\left(s^*\right)+D''\left(t-s^*\right)\right)^3} \tag{3}$$

**Lemma 9** *For $t \geq \bar{s}$ :*

1. *Conformity is locally weakly decreasing in $t$ if and only if $\frac{ds^*}{dt} \geq 0$.*
2. *In corner solutions, relative concession is locally constant. In inner solutions, relative concession is locally weakly increasing in $t$ if and only if $\left(s^*-\bar{s}\right)P''\left(s^*-\bar{s}\right) \geq \left(t-s^*\right)D''\left(t-s^*\right)$.*

***Proof.*** *1) trivially follows from Definition 1. 2) In corner solutions $s^*\left(t\right) \in \left\{\bar{s}, t\right\}$ which implies that, locally, relative concession is either equal to 1 or equal to 0. For inner solutions: By differentiating the expression (in Definition 2) for relative concession w.r.t. $t$, performing a*

*few algebraic steps making use of equality of the first derivative in inner solutions and equations 2 and 3, it can be verified that the derivative is proportional to* $\frac{(s^*-\bar{s})P''(s^*-\bar{s})-(t-s^*)D''(t-s^*)}{P''(s^*-\bar{s})+D''(t-s^*)}$. *In min points the denominator is positive and the inequality then follows.* ■

### A.1.2 The possible locations of the norm in equilibrium

**Lemma 10** *Suppose* $\bar{s}$ *is the average stance in society and* $t \sim U(t_l, t_h)$. *Then, for any positive* $\alpha$ *and* $\beta$ *there is an equilibrium where* $\bar{s} = \frac{t_l+t_h}{2}$.
***Proof.*** *Let* $d \equiv \min\{t_h - \bar{s}, \bar{s} - t_l\}$. *Since the solution for any type's optimization problem depends only on the distance from* $\bar{s}$, *we know that the distribution of the stances of all the types in the range* $[\bar{s}-d, \bar{s}+d]$ *is symmetric around* $\bar{s}$. *Thus* $\bar{s}$ *is the average stance for this range of types. If* $\bar{s} = \frac{t_l+t_h}{2}$, *then* $[\bar{s}-d, \bar{s}+d] = [t_l, t_h]$, *and so* $\bar{s}$ *is the average stance for all types in society. It thus follows that* $\bar{s} = \frac{t_l+t_h}{2}$ *can be sustained as a social norm in equilibrium for any values positive of* $\alpha$ *and* $\beta$. ■

**Lemma 11** *Suppose* $\bar{s}$ *is the average stance in society and* $t \sim U(t_l, t_h)$. *Let* $d \equiv \min\{t_h - \bar{s}, \bar{s} - t_l\}$. *Then* $\bar{s} \neq \frac{t_l+t_h}{2}$ *can be sustained in equilibrium only if* $s^*(t) = \bar{s}\ \forall t \in [t_l, t_h] \setminus [\bar{s}-d, \bar{s}+d]$.
***Proof.*** *Since the solution for any type's optimization problem depends only on the distance from* $\bar{s}$, *we know that the distribution of the stances of all the types in the range* $[\bar{s}-d, \bar{s}+d]$ *is symmetric around* $\bar{s}$. *Thus* $\bar{s}$ *is the average stance for this range of types. If* $\bar{s} > \frac{t_l+t_h}{2}$, *then by definition* $\bar{s}+d = t_h$, *and so* $\forall t \in [t_l, t_h] \setminus [\bar{s}-d, \bar{s}+d] = [t_l, \bar{s}-d]$ *we have* $s^*(t) \leq \bar{s}$. *For* $\bar{s}$ *to be the average of all stances, it is then necessary that* $s^*(t) = \bar{s}\ \forall t < \bar{s} - d$. ■

### A.1.3 Transformation from individually chosen stances to the distribution of stances

We now analyze the density function of the chosen stances in society ($PDF$). We restrict ourselves to cases where the optimal stance of each type is uniquely determined.[24] We divide the range of types into $n+1$ subranges

$$T_0 = [t_{low}, t_1], T_1 = [t_1, t_2], ...T_n = [t_n, t_{high}],$$

such that:

1. In each subrange, the function $s^*(t)$ either consists of only corner solutions or consists of only inner solutions.

[24] Otherwise we have no way of determining the chosen stance of some types.

2. In case of corner solutions we have either $s^*(t) = t \ \forall t \in T_{i.}$ or $s^*(t) = \bar{s} \ \forall t \in T_{i.}$

3. In case of inner solutions, $s^*(t)$ is continuous and strictly monotonic in a subrange.

We now investigate separately the contribution of each such subrange of types to the resultant $PDF$. The contribution of each such part is called a *partial* $PDF$, to be denoted $pPDF_{T_i}$, where

$$PDF = \sum_i pPDF_{T_i}.$$

**Inner solutions** Here we investigate the properties of the $pPDF_{T_i}$ (dropping the $T_i$ index where possible) in subranges with inner solutions. Denote by $s^*_{\min}$ the lowest stance taken by a type in the subrange (strict monotonicity ensures that this type is unique). Let $M_i(\tilde{s}^*)$ be the mass of types in $T_i$ with stances in the range $(s^*_{\min}, \tilde{s}^*]$ for some $\tilde{s}^*$:

$$M_i(\tilde{s}^*) \equiv \int_{s^*_{\min}}^{\tilde{s}^*} pPDF_{T_i} ds = \begin{cases} \int_{t_i}^{t(\tilde{s}^*)} f(\tau)\, d\tau \text{ if } s^*(t) \text{ is increasing in the subrange } T_i \\ \int_{t(\tilde{s}^*)}^{t_{i+1}} f(\tau)\, d\tau \text{ if } s^*(t) \text{ is decreasing in the subrange } T_i \end{cases}$$

where $t(\tilde{s}^*) \equiv \{t \text{ s.t. } s^*(t) = \tilde{s}^*\}$ and $f(t)$ is the density function of $t$.

If the distribution of types is uniform, i.e. $f(t) = 1/(t_h - t_l)$, we get:

$$M_i(\tilde{s}^*) = \begin{cases} \frac{t(\tilde{s}^*) - t_i}{t_h - t_l} \text{ if } s^*(t) \text{ is increasing in the subrange } T_i \\ \frac{t_{i+1} - t(\tilde{s}^*)}{t_h - t_l} \text{ if } s^*(t) \text{ is decreasing in the subrange } T_i \end{cases} \quad (4)$$

$$pPDF_{T_i}(\tilde{s}^*) = \frac{dM_i(\tilde{s}^*)}{d\tilde{s}^*} = \frac{1}{t_h - t_l} \left| \frac{dt}{ds^*}|_{\tilde{s}^*} \right| \quad (5)$$

Note that the last derivation is valid only if $\frac{ds^*}{dt}|_{\tilde{s}^*} \neq 0$ as otherwise $\frac{dt}{ds^*}$ is not defined. This is ensured under the strict monotonicity of $s^*(t)$. We then have, by using the implicit function theorem twice:

$$\frac{d(pPDF(\tilde{s}^*))}{ds^*} = \begin{cases} \frac{1}{t_h - t_l} \frac{d^2t}{ds^{*2}}|_{\tilde{s}^*} \text{ if } \frac{dt}{ds^*}|_{\tilde{s}^*} > 0 \\ -\frac{1}{t_h - t_l} \frac{d^2t}{ds^{*2}}|_{\tilde{s}^*} \text{if } \frac{dt}{ds^*}|_{\tilde{s}^*} < 0 \end{cases}. \quad (6)$$

In inner solutions, the following result then applies.[25]

**Lemma 12** *In inner solutions, the pPDF is locally strictly increasing at $s^*$ if $\frac{d^2s^*}{dt^2}$ is negative, and strictly decreasing at $s^*$ if $\frac{d^2s^*}{dt^2}$ is positive.*
***Proof.*** *From equation 6, it follows that the pPDF is increasing if $\frac{dt}{ds^*}$ and $\frac{d^2t}{ds^{*2}}$ have the same sign and decreasing if $\frac{dt}{ds^*}$ and $\frac{d^2t}{ds^{*2}}$ have opposite signs. We then use the fact that $\frac{d^2s^*}{dt^2} < 0$ if $\frac{dt}{ds^*}$ and $\frac{d^2t}{ds^{*2}}$ have the same sign, and $\frac{d^2s^*}{dt^2} > 0$ if $\frac{dt}{ds^*}$ and $\frac{d^2t}{ds^{*2}}$ have opposite signs.* ■

**Corner solutions.** There are two candidate corner solutions. The first is $s^*(t) = t$. In a subrange of these corner solutions, the $pPDF$ is simply a uniform distribution with the trivial properties

$$pPDF(\tilde{s}^*) = \frac{1}{t_h - t_l}\frac{dt}{ds^*}|_{\tilde{s}^*} = \begin{cases} \frac{1}{t_h - t_l} \text{ if } \tilde{s}^*(t) = t \\ 0 \text{ otherwise} \end{cases}$$
$$\frac{d(pPDF)}{ds} = 0.$$

The other candidate corner solution is $s^*(t) = \bar{s}$. The solution of this equation is independent of $t$, so in a subrange of these corner solutions, the $pPDF$ is a degenerate single peak with a mass equalling the mass of types within that subrange.

$$pPDF_{T_i}(\tilde{s}^*) = \begin{cases} \frac{t_{i+1} - t_i}{t_h - t_l} \text{ if } s^* = \bar{s} \\ 0 \text{ otherwise} \end{cases}$$

## A.2 Proof of Proposition 1

For strict societies, the proofs of the first three statements are contained in the proof to part (1) of Proposition 3. As for the fourth statement, when $\bar{s} \in [t_h - \Delta, t_l + \Delta]$ we get that $\max\{t_h - \bar{s}, \bar{s} - t_l\} \leq \Delta$, hence $|t - \bar{s}| < \Delta$ for every type. In the proof to part 1 of Proposition 3 we show that when $\beta < \alpha$, $s^*(t) = \bar{s}$ for every $t$ such that $|t - \bar{s}| < \Delta$. It thus follows that $s^*(t) = \bar{s}$ for all types, and so $\bar{s}$ is the average of all stances, as required to constitute a norm. Lemma 10 states that $\bar{s} = \frac{t_l + t_h}{2}$ can also be sustained in equilibrium, which concludes the proof.

In liberal societies we have $1 = \alpha < \beta$. Solving for the range $t > \bar{s}$ and then using symmetry around $\bar{s}$, it is easy to verify that types sufficiently close to the norm declare their type, while types sufficiently far from the

[25] Note that the previous expressions capture the "local" contribution to the $PDF$. E.g., there can be cases where a stance $s$ is chosen (as a corner solution) by the type $t = s$ and at the same time (as an inner solution) by a different type with $t > s$. In such a case these two types will belong to two separate subranges ($T_i$) hence will contribute to two separate $pPDF's$.

norm have an inner solution $s$ s.t. $P'(|s-\bar{s}|) = 1$ $(= D')$. This proves statement (1). In the subrange where all declare their type, relative concession equals 0, and in the subrange where all declare the same inner stance it is increasing, as follows from Lemma 9 part (2) with $P'' > 0$ and $D'' = 0$ (use symmetry to see that it holds also for the range $t < \bar{s}$). This proves statement (2). Before proving statement (3) we prove the fourth statement. By Lemma 10 we know that $\bar{s} = \frac{t_l+t_h}{2}$ can be sustained in equilibrium. Thus, we only need to show that $\bar{s} \neq \frac{t_l+t_h}{2}$ cannot be an equilibrium. This follows directly from Lemma 11 and the fact that no one chooses $s^*(t) = \bar{s}$. Finally, to see statement (3), note that the value of $s$ s.t. $P'(|s-\bar{s}|) = 1$ is given by solving $\beta K (s-\bar{s})^{\beta-1} = 1$, which yields $s - \bar{s} = (\beta K)^{\frac{1}{1-\beta}}$. Given that $\bar{s} = \frac{t_l+t_h}{2}$, we get that if $t_h - t_l > 2(\beta K)^{\frac{1}{1-\beta}}$ then types at both edges of the distribution choose the inner solutions $s^* = \bar{s} \pm (\beta K)^{\frac{1}{1-\beta}}$. These types form the modes of a bimodal distribution while types closer to the norm form a uniform part. If the range of types is smaller than $2(\beta K)^{\frac{1}{1-\beta}}$ all types choose $s^*(t) = t$, implying a uniform distribution.■

## A.3 Proof of Proposition 2

Since the functions are symmetric around $\bar{s}$, we present only the proof for the range of $t \geq \bar{s}$.

**Parts 1 and 2:**

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

$$-\alpha (t-s)^{\alpha-1} + \beta K (s-\bar{s})^{\beta-1} = 0 \tag{7}$$

$$(\alpha-1)\alpha (t-s)^{\alpha-2} + (\beta-1)\beta K (s-\bar{s})^{\beta-2} > 0 \tag{8}$$

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

$\alpha, \beta > 1$: That every $t$ has a unique inner solution can be easily verified using equations (7) and (8). The statements that $|s^*(t) - \bar{s}|$ is increasing either convexly or concavely follow from applying the implicit function theorem twice to equation (7) to get $ds^*/dt$ and $d^2s^*/dt^2$. Since all types have inner solutions, the statements regarding relative concession follow from restating the inequality in part 2 of Lemma 9 explicitly for power functions, and substituting the FOC into it. Plugging the expressions for the derivatives of $P$ and $D$ into equation (3), we get that $\frac{d^2s^*}{dt^2} > 0$ when $\alpha > \beta$ and $\frac{d^2s^*}{dt^2} < 0$ when $1 < \alpha < \beta$. Using the derived expression for $d^2s^*/dt^2$ it then follows from Lemma 12 that the $pPDF$ is decreasing in the distance to $\bar{s}$ when $\alpha > \beta$ and increasing

when $1 < \alpha < \beta$. As $s^*(t)$ is monotonic, the $pPDF$ represents the total $PDF$. From the symmetry of the functions around $\bar{s}$ (and the central location of the norm, as implied by Lemma 11 when no type chooses $s^*(t) = \bar{s}$), it then follows that the distribution is unimodal when $\alpha > \beta$ and bimodal when $\alpha < \beta$. Finally, the convexity of $P$ and $D$ implies that $\forall t \geq \bar{s}$ we have $0 \leq \frac{ds^*}{dt} = \frac{D''(t-s^*)}{P''(s^*)+D''(t-s^*)} \leq 1$. Hence, it follows from part 1 of Lemma 9 that conformity is decreasing $\forall t \geq \bar{s}$.

$1 = \beta < \alpha$: It is easy to verify that types sufficiently close to the norm declare the norm (this is true for any $K > 0$) and types sufficiently far from the norm have a unique inner solution. For the subrange where all declare the norm $ds^*/dt = 0$ and hence $d^2s^*/dt^2 = 0$. For the subrange with inner solutions using $\beta = 1$ and $\alpha > 1$ in equation (2) implies $ds^*/dt = 1$ and hence $d^2s^*/dt^2 = 0$. Applying these results to Lemma 9 yields the first three statements of part 1. Since for any $K > 0$ there exist some types sufficiently close to the norm who declare the norm, there will always be a peak at the norm. Since in the other subrange $ds^*/dt = 1$ this implies a unimodal distribution in total.

**Part 3:**

For the case $1 = \alpha < \beta$, the proofs for all the statements are contained in the proof of Proposition 1, except for (i) the statement about conformity, which follows from the fact that types sufficiently close to the norm declare their type and types sufficiently far from the norm at each side of it declare the same stance $s = \bar{s} \pm (\beta K)^{\frac{1}{1-\beta}}$ (implying that $|s^*(t) - \bar{s}|$ is first increasing and then it is constant); (ii) showing that $t_h - t_l > 2\Delta$ is a sufficient condition for a bimodal distribution of $s^*$, which follows from the fact that to get bimodality it was shown that $t_h - t_l$ should be greater than $2(\beta K)^{\frac{1}{1-\beta}}$, and noting that $2(\beta K)^{\frac{1}{1-\beta}} < 2K^{\frac{1}{1-\beta}} = 2\Delta$.

For the case $\alpha < 1 < \beta$, we will show that if $t_h - t_l > 2\Delta$, then a) $|s^*(t) - \bar{s}|$ is first increasing then decreasing in $|t - \bar{s}|$, implying non-monotonic conformity; b) the relative concession is increasing in $|t - \bar{s}|$; and c) if, furthermore, $t \sim U(t_l, t_h)$, then the distribution of $s^*$ is bimodal.

a) We will first show that the only relevant corner solution is $s^* = t$, then that types close to the norm choose this corner solution. In order to find the global minimum we first need to investigate the behavior of $L(s,t)$ near the corner solutions.

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

Hence $L'(\bar{s},t) < 0$ and $L'(t,t) < 0$ since $\alpha < 1$. Therefore $s = t$ may be a solution to the minimization problem while $s = \bar{s}$ will not. The

candidate solution $s = t$ will now be compared to potential local minima in the range $[\overline{s}, t]$. In inner solutions $L'(s,t) = 0$ and hence we get

$$\alpha (t-s)^{\alpha-1} = \beta K (s-\overline{s})^{\beta-1} \Rightarrow (t-s)^{\alpha-1} (s-\overline{s})^{1-\beta} = K\beta/\alpha$$

Define $f(s) \equiv (t-s)^{\alpha-1} (s-\overline{s})^{1-\beta}$. For the existence of an inner min point it is necessary that $f(s) = \beta K/\alpha$ for some $s \in ]\overline{s}, t[$. Notice that $f(s)$ is strictly positive in $]\overline{s}, t[$, and that $f(s) \to \infty$ at both edges of the range (i.e. at $s = \overline{s}$ and at $s = t$). This means that $f(s)$ has at least one local minimum in $]\overline{s}, t[$. We now proceed to check whether this local minimum is unique:

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

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

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

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

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

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

Finally we show that if there exists any type $t_0$ who chooses the inner solution, then all types with $t > t_0$ have an inner solution too. Then we

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

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

and since $\alpha < 1 \leq \beta$, for small enough $|t-\overline{s}|$ the dominant element is $(t-\tilde{s})^{\alpha}$ and so $\Delta L$ is negative (i.e., types close to the norm choose the corner solution of $s^* = t$). Differentiating $\Delta L$ with respect to $t$ yields

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

Using the first order condition

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

Differentiating once more

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

By equation 2 we have that $\frac{d\tilde{s}}{dt} < 0$ in an inner solution when $D$ is concave, and so $\Delta L''_t > 0$. Hence $\Delta L$ is strictly increasing and strictly convex, implying that for a broad enough range of types, types sufficiently far from the norm have an inner solution. Moreover, at this subrange of types, $\frac{ds^*}{dt} < 0$. This implies that $s^*$ is first increasing (in the subrange of types with $s^* = t$), and then decreasing (in the subrange of types with inner solutions).

b) By the definition of relative concession it equals 0 at the subrange of types choosing $s^* = \overline{t}$, and then it rises at the cutoff where $\Delta L = 0$,

and keeps rising as $t$ increases (this follows from restating the inequality in part 2 of Lemma 9 explicitly for power functions, and substituting the FOC in it).

c) This implies that if the range of types is narrow, all types state their type, creating a uniform distribution of stances. If the range of types is broad enough to include types with inner solutions, then on top of the uniform part there is a peak on each side of $\bar{s}$ (since $\bar{s}$ must equal $\frac{t_l+t_h}{2}$, as implied by Lemma 11 when no type chooses $s^*(t)=\bar{s}$). These peaks are inside the uniform distribution. To see this, note that for the type $t$ who is just indifferent between the corner and inner solution, the inner solution would entail $s^*(t) \leq t$. Together with the previous result that $\frac{ds^*}{dt} < 0$ we get that all types with inner solutions choose statements within the bounds of the uniform part.

To see the shape of the distribution of stances, note first that $\frac{dt}{ds^*} < 0$ because $\frac{ds^*}{dt} < 0$. As for $\frac{d^2t}{ds^{*2}}$, we have:

$$\frac{d^2t}{ds^{*2}} = \frac{d}{ds^*}\left(\frac{dt}{ds^*}\right) = \left(\frac{\beta K}{\alpha}\right)^{\frac{1}{\alpha-1}} \frac{\beta-1}{\alpha-1}\left(\frac{\beta-1}{\alpha-1}-1\right)\left(s^*-\bar{s}\right)^{\frac{\beta-1}{\alpha-1}-2}.$$

Substituting $t-s^* = \left(\frac{\beta K}{\alpha}\right)^{\frac{1}{\alpha-1}}\left(s^*-\bar{s}\right)^{\frac{\beta-1}{\alpha-1}}$ in this expression we get that

$$\frac{d^2t}{ds^{*2}} = \frac{(\beta-1)(t-s^*)}{(\alpha-1)(s^*-\bar{s})^2}\left(\frac{\beta-1}{\alpha-1}-1\right).$$

Since both $\frac{(\beta-1)(t-s^*)}{(\alpha-1)(s^*-\bar{s})^2}$ and $\left(\frac{\beta-1}{\alpha-1}-1=\frac{\beta-\alpha}{\alpha-1}\right)$ are negative, we get that $\frac{d^2t}{ds^{*2}} > 0$, which together with $\frac{dt}{ds^*} < 0$ implies by the inverse function theorem that $\frac{d^2s^*}{dt^2} > 0$. Thus by Lemma 12 the $pPDF$ for the inner solutions is decreasing towards the edges.■

## A.4 Proof of Proposition 3

**Parts 1 and 2**

The second-order condition (equation 8) is positive when $\alpha < \beta \leq 1$ or $\beta < \alpha \leq 1$, which implies that any inner extreme point is a maximum. The corner solutions are then either $L(s=\bar{s}) = |t-\bar{s}|^\alpha$ or $L(s=t) = K|t-\bar{s}|^\beta$. When $\beta < \alpha$ this implies that $L(s=\bar{s}) < L(s=t)$ iff $|t-\bar{s}| < \Delta = K^{\frac{1}{\alpha-\beta}}$, and so $s^*(t) = t$ iff $|t-\bar{s}| \geq \Delta$, and $s^*(t) = \bar{s}$ iff $|t-\bar{s}| < \Delta$. This means that conformity is initially constant and then strictly decreasing, which altogether implies that conformity is everywhere weakly decreasing in $|t-\bar{s}|$. When $\alpha < \beta$ the converse holds, with $s^*(t) = t$ iff $|t-\bar{s}| < \Delta$, and $s^*(t) = \bar{s}$ iff $|t-\bar{s}| \geq \Delta$, which means that conformity is initially decreasing in $t$ but then sharply increases to

full conformity at $|t-\bar{s}|=\Delta$ (where it also stays). In the segment of types choosing $s^*(t)=\bar{s}$, the relative concession is equal to 1, while in the segment of types choosing $s^*(t)=t$, the relative concession is 0. From this, it follows that the relative concession is weakly decreasing with the distance to $\bar{s}$ for $\beta<\alpha$ and weakly increasing for $\alpha<\beta$. As for the distribution of $s^*$, we start with the case of $\beta<\alpha$. First note that a broad enough range of types, $t_h-t_l>2\Delta$, implies that $s^*(t)\neq\bar{s}$ for the type furthest away from $\bar{s}$, which by Lemmas 10 and 11 implies that $\bar{s}=\frac{t_l+t_h}{2}$ is the only possible norm in equilibrium. The individual choices will then lead to a distribution of stances that consists of a peak at $\bar{s}$ and two equally sized uniform tails at the extreme ends of the distribution, detached from the peak. In the case of $\alpha<\beta$, the sufficient condition for having a peak at $\bar{s}$ is to have types with $|t-\bar{s}|>\Delta$ at one side of $\bar{s}$, which holds when $t_h-t_l>2\Delta$.

**Part 3**

We perform the proof for $t\geq\bar{s}$. The opposite case is similar. We will prove that if $t_h-t_l>2\Delta$, then: a) types close enough to the norm fully conform, while types far from the norm choose an inner solution and $|s^*(t)-\bar{s}|$ is increasing for them; b) conformity is weakly decreasing in $|t-\bar{s}|$; c) the relative concession is weakly decreasing in $|t-\bar{s}|$; and d) if furthermore, $t\sim U(t_l,t_h)$, then the distribution is discontinuously trimodal with a central peak at $\bar{s}$ and a detached uniform part on each side, peaking at the edge of the range. Along the way we will also show that for a sufficiently narrow range of types, the distribution is degenerate at $\bar{s}$.

a) We will first show that the only relevant corner solution is $s^*=\bar{s}$. In order to find the global minimum we first need to investigate the behavior of $L(s,t)$ at the edges of this range.

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

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

$$\alpha(t-s)^{\alpha-1}=\beta K(s-\overline{s})^{\beta-1}\Rightarrow(t-s)^{\alpha-1}(s-\overline{s})^{1-\beta}=\beta K/\alpha.$$

Define $f(s)\equiv(t-s)^{\alpha-1}(s-\overline{s})^{1-\beta}$. For the existence of an inner min point it is necessary that $f(s)=\beta K/\alpha$ for some $s\in]\overline{s},t[$. Notice that $f(s)$ is strictly positive in $]\overline{s},t[$, and that $f(s)=0$ at both edges of the

range (i.e. at $s = \overline{s}$ and at $s = t$). This means that $f(s)$ has at least one local maximum in $]\overline{s}, t[$. We now proceed to check whether this local maximum is unique:

$$f'(s) = (t-s)^{\alpha-2}(s-\overline{s})^{-\beta}\left[(1-\beta)(t-s) - (\alpha-1)(s-\overline{s})\right]$$

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

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

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

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

Finally we show that if there exists any type $t_0$ who chooses the inner solution then all types with $t > t_0$ have an inner solution. Then we show that types close enough to the norm fully conform, and that in the range of inner solutions $|s^*(t) - \overline{s}|$ is increasing in $t$. First notice that $f(s)$ is increasing in $t$, so if there exists a local minimum of $L(s, t_0)$ for some $t_0$, then there exists a local minimum of $L(s,t)$ for $t > t_0$ too. Also note that $f(s)$ is increasing in $t$ with $\lim_{t\to\infty} f(s) = \infty > \beta K/\alpha$ (for $s \in ]\overline{s}, t[$), implying an inner local min point exists for a broad enough range of types. Second, if there is an inner solution to the minimization problem for some $t_0$ then there is also an inner solution to the minimization problem for $t > t_0$. To see this let $\Delta L \equiv L(\overline{s}, t) - L(\tilde{s}, t)$, where $\tilde{s}$ is

the stance at which $L(s,t)$ gets the local minimum. Type $t$ prefers the inner solution to the corner solution if and only if $\Delta L$ is positive. Thus we need to show that $\Delta L$ is negative for small enough $|t-\bar{s}|$ but is increasing in $t$ (and so if $\Delta L$ is positive for $t_0$ it is positive for $t_1$ too).

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

and since $\beta \leq 1 < \alpha$, for small enough $|t-\bar{s}|$ the dominant element is $K\left(\tilde{s}-\overline{s}\right)^{\beta}$ and so $\Delta L$ is negative (i.e., types close to the norm choose the corner solution of $s^* = \overline{s}$). Differentiating $\Delta L$ with respect to $t$ yields

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

Using the first order condition

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

Differentiating once more

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

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

b) Looking at $|s^*(t)-\bar{s}|$ as $|t-\bar{s}|$ gradually increases, we first have $|s^*(t)-\bar{s}|$ constant and equal to 0 in the range where $s^*(t) = s^*$. Then at some point we reach the first type who chooses an inner solution, and so for her $|s^*(t)-\bar{s}| > 0$, and afterwards $|s^*(t)-\bar{s}|$ was shown to keep increasing. Therefore conformity is everywhere weakly decreasing in $|t-\bar{s}|$.

c) By the definition of relative concession it equals 1 at the range of types choosing $s^* = \overline{s}$, and then it falls at the cutoff where $\Delta L = 0$, and keeps falling as $t$ increases (this follows from restating the inequality in part 2 of Lemma 9 explicitly for power functions, and substituting the FOC in it).

d) If the range of types is narrow, all types state the norm, hence follows a degenerate distribution at $\overline{s}$. Otherwise, if the range of types

is broad enough (so that some have an inner solution), the resulting distribution of stances has a peak at $s = \bar{s}$ with a tail at each side of it. The tails will be of equal size because $s^*(t) \neq \bar{s}$ for the type furthest away from $\bar{s}$, which by Lemmas 10 and 11 implies that $\bar{s} = \frac{t_l + t_h}{2}$, and so the distribution of stances is symmetric. The tails are detached, since for the type $t$ who is indifferent to either the corner or the inner solution the inner solution $s^*$ is necessarily strictly greater than $\bar{s}$ (because $L'(\bar{s}, t) = \infty$ while $L'(s^*, t) = 0$ in inner solutions).

For the shape of these tails, note first that $\frac{dt}{ds^*} > 0$ because $\frac{ds^*}{dt} > 0$. Moreover, $\frac{d^2t}{ds^{*2}} = \frac{(\beta-1)(t-s^*)}{(\alpha-1)(s^*-\bar{s})^2}\left(\frac{\beta-1}{\alpha-1} - 1\right)$ (see the proof of Proposition 2 part 3). Since both $\frac{(\beta-1)(t-s^*)}{(\alpha-1)(s^*-\bar{s})^2}$ and $\left(\frac{\beta-1}{\alpha-1} - 1\right)$ are negative, we get that $\frac{d^2t}{ds^{*2}} > 0$, which together with $\frac{dt}{ds^*} > 0$ implies by the inverse function theorem that $\frac{d^2s^*}{dt^2} < 0$. Thus by Lemma 12 the $pPDF$ for the inner solutions is increasing towards the edges, which together with the peak at $\bar{s}$ implies a trimodal distribution of stances.■

## A.5 Proof of Proposition 4

Lemma 10 ensures that $\bar{s} = \frac{t_l + t_h}{2}$ is a possible equilibrium. Furthermore:

1) From parts (1) and (3) of Proposition 3 we know that if $\alpha \leq 1$ then types with $|t - \bar{s}| \leq \Delta$ fully conform and if $1 < \alpha$ then there exists a cutoff distance $\delta < \Delta$ such that types with $|t - \bar{s}| \leq \delta$ fully conform. So if $\alpha \leq 1$ and $t_h - t_l < 2\Delta$ or alternatively $1 < \alpha$ and $t_h - t_l < 2\delta$, and if $\left|\bar{s} - \frac{t_l + t_h}{2}\right|$ is small enough to ensure that the type furthest away from $\bar{s}$ is still within distance $\Delta$ (in the case $\alpha \leq 1$) or $\delta$ (in the case $1 < \alpha$) from $\bar{s}$, then $s^*(t) = \bar{s}\ \forall t$. In this case $\bar{s}$ is obviously the average of all stances and this concludes sufficiency. To show necessity, we need to show that if $\alpha \leq 1$ and $t_h - t_l \geq 2\Delta$ or alternatively $1 < \alpha$ and $t_h - t_l \geq 2\delta$, there exist no equilibria with $\bar{s} \neq \frac{t_h + t_l}{2}$. To see this, note that $\bar{s} \neq \frac{t_h + t_l}{2}$ would imply that $|t - \bar{s}| > \Delta$ (in the case $\alpha \leq 1$) or $|t - \bar{s}| > \delta$ (in the case $1 < \alpha$) for the type furthest away from $\bar{s}$. Then, parts (1) and (3) of Proposition 3 imply that $s^*(t) \neq \bar{s}$ for that type, and by Lemma 11 $\bar{s} \neq \frac{t_h + t_l}{2}$ cannot be sustained in equilibrium.

2) From Proposition 3 part (2) we know that types with $|t - \bar{s}| < \Delta$ choose $s^* = t$, while types with $|t - \bar{s}| > \Delta$ choose $\bar{s}$ as their stance (and therefore do not affect its location). If $t_h - t_l > 2\Delta$, take any $\bar{s} \in [t_l + \Delta, t_h - \Delta]$. This implies that $\bar{s} - t_l > \Delta$ and $t_h - \bar{s} > \Delta$, and so ensures that the whole uniform section of the stance distribution, $[\bar{s} - \Delta, \bar{s} + \Delta]$, is contained within $[t_l, t_h]$, with $\bar{s}$ located at the center of the uniform section, implying that it is the average stance. This concludes sufficiency. Otherwise, suppose $t_h - t_l < 2\Delta$. Then $[t_l + \Delta, t_h - \Delta]$ is an empty set, therefore either $\bar{s} > t_h - \Delta$ or $\bar{s} < t_l + \Delta$ (or both).

Suppose $\bar{s} > t_h - \Delta$. Then $d \equiv t_h - \bar{s} < \Delta$, and so $s^*(t) = t \neq \bar{s}$ for types with $t \in [\bar{s} - \Delta, \bar{s} - d]$, which by Lemma 11 implies that $\bar{s}$ cannot be sustained in equilibrium. A corresponding argument applies to $\bar{s} < t_l + \Delta$. Repeating the same exercise for the case $t_h - t_l = 2\Delta$ implies that $\bar{s} = \frac{t_l + t_h}{2}$ is the unique possible norm in this case. This concludes necessity.■

## A.6 Proof of Proposition 5

$s_h = \bar{s} + 1$ implies that $P$ in the range $]\bar{s}, \bar{s} + 1[$ strictly decreases in $\beta$ while $P$ at $\{\bar{s}, \bar{s} + 1\}$ is independent of $\beta$. Hence, a decrease of $\beta$ implies that any stance in $]\bar{s}, \bar{s} + 1[$ becomes less attractive, while the attractiveness of the stances $\bar{s}$ and $\bar{s} + 1$, where pressure is fixed at 0 and $K$ respectively, stays the same. Thus, for any given $t$, if $s^*(t) \in \{\bar{s}, \bar{s} + 1\}$ for $\beta$ then $s^*(t) \in \{\bar{s}, \bar{s} + 1\}$ also for any $\beta' < \beta$.

## A.7 Proof of Proposition 7

At any stance $s > \bar{s}$, the pressure $P = K(s - \bar{s})^\beta$ strictly decreases in $\beta$ if $s \in ]\bar{s}, \bar{s} + 1[$ and strictly increases in $\beta$ if $s > \bar{s} + 1$, while staying constant at $s = \bar{s}$ and $s = \bar{s} + 1$. Therefore, for any $t > \bar{s}$, the function $L(s, t)$ (which $t$ minimizes at $[\bar{s}, t]$ to get $s^*(t)$) decreases in $\beta$ at the range $]\bar{s}, \bar{s} + 1]$ and increases in $\beta$ at the range $s > \bar{s} + 1$. Considering now the effect of increasing $\beta$, this implies that if before the increase $s^*(t) \in S_{low} = ]\bar{s}, \bar{s} + 1]$, then $s^*(t) \in S_{low}$ also after the increase, and if $s^*(t) \in S_{high}$ after the increase, then $s^*(t) \in S_{high}$ also before it. Therefore, $|S_{low}|$ is weakly greater after the increase, while $|S_{high}|$ is weakly smaller after the increase. Together with the fact that $s^*(t) \leq \bar{s}$ for types with $t \leq \bar{s}$ (and so these types do not affect $|S_{low}|$ and $|S_{high}|$), we get that $|S_{low}| - |S_{high}|$ increases in $\beta$.■

## A.8 Proof of Proposition 6

If $\beta > 1$, then no one with $t \neq \bar{s}$ chooses $\bar{s}$ as her stance regardless of the value of $K$, as a small deviation toward $t$ would reduce $D$ without affecting $P$. Thus, increasing $K$ will not increase the number of individuals stating $\bar{s}$. Suppose, however, that $\beta \leq 1$, and consider an increase in $K$ from some $K_1$ to some $K_2 > K_1$. An increase in $K$ implies that the loss in any possible stance except for $\bar{s}$ increases, so anyone who chooses $\bar{s}$ under $K_1$ also chooses it under $K_2$. This implies that the number of people stating $\bar{s}$ will not decrease when $K$ increases. Furthermore, if $t_h - t_l > 2\Delta$, then under $K_1$ there are types who are indifferent between choosing $\bar{s}$ and choosing some different stance (see Proposition 3). These types will strictly prefer $\bar{s}$ under $K_2$, implying a strict increase in the number of people stating $\bar{s}$ as $K$ increases from $K_1$ to $K_2$.

## A.9 Proof of Corollary 8

Follows directly from propositions 2 and 3.

# B Appendix: data and empirical details for Section 7.1

The data was accessed in February 2015 from www.worldvaluessurvey.org through their online analysis toolbox.

- **Countries**: Algeria, Azerbaijan, Palestine, Iraq, Kazakhstan, Jordan, Kyrgyzstan, Lebanon, Libya, Malaysia, Morocco, Pakistan, Tunisia, Turkey, Uzbekistan, Yemen. 16 countries in total. Each question is answered by around 1000 individuals in each country. There are a few additional Muslim countries but they do not have data on the praying variable.

- **$\beta$ proxy variable:** The proxy for $\beta$ was constructed using answers to question V154 measuring agreement with the statement "The only acceptable religion is my religion". The measure was constructed by taking the share strongly disagreeing minus the share strongly agreeing. This implies a high value will be the equivalent of a high $\beta$. Note that values can be negative. Plain "Agree" and "Disagree" were not used since it would be unclear how the ordinality in-between answers could be collapsed to a single measure (e.g., if 50% answer agree and 50% disagree, is that society more or less strict than if 50% answer strongly agree and 50% answer strongly disagree?).

- **Religiosity variable (t-distribution)**: The share answering "A religious person" to question V147. The other possible answers are "Not a religious person", "Atheist", "No answer" and "Don't know".

- **Harshness variable (K):** Answers to question V132 "Democracy: Religious authorities interpret the laws". The possible answers range from 1) "Not an essential characteristic of democracy" to 10) "An essential characteristic of democracy". We used the mean.

- **Praying variable (s)**: V146, "How often do you pray?". Apart from "Don't know" and "No Answer" the alternatives were 1) "Several times a day", 2) "Once a day", 3) "Several times each week", 4) "Only when attending religious services", 5) "Only on special holy days", 6) "Once a year", 7) "Less often than once a

year" and 8) "Never, practically never". In the regression table, the total share answering either 1 or 8 is defined as "Extremes". The share answering 2 or 3 is defined as "Religious non-extremes". The share answering 4 or 5 is defined as "Mid". The share answering 6 or 7 is defined as "Secular non-extremes".

Table 1: Summary statistics

| Variable | Mean | Min | Max |
|---|---|---|---|
| $\beta$ proxy | -45.2 | -81.4 (Libya) | 14.5 (Kazakhstan) |
| Religiousity % | 71.0 | 26.7 (Azerbaijan) | 99.7 (Pakistan) |
| Harshness | 5.3 | 3.69 (Azerbaijan) | 6.42 (Yemen) |
| Praying: Extremes % | 69.0 | 43.6 (Lebanon) | 93.6 (Tunisia) |
| Praying: Religious non-extremes % | 14.8 | 1.2 (Tunisia) | 41.9 (Lebanon) |
| Praying: Mid % | 9.6 | 1.5 (Jordan) | 34.4 (Azerbaijan) |
| Praying: Secular non-extremes % | 5.4 | 1 (Jordan) | 13.9 (Uzbekistan) |
| Praying: Several times per day % | 53.2 | 6.3 (Kazakhstan) | 86.4 (Jordan) |
| Pracying: Never, practically Never % | 15.8 | 0.6 (Pakistan) | 53.7 (Uzbekistan) |

Table 2: Regression results

| | Extremes | Extremes | Religious non-extremes | Religious non-extremes | Mid | Mid | Secular non-extremes | Secular non-extremes |
|---|---|---|---|---|---|---|---|---|
| Intercept | 45.5*** | 57.2*** | 23.1*** | 2.98 | 23.2*** | 38.5*** | 9.0*** | 3.5 |
| $\beta$ proxy | -0.53*** | -0.58*** | 0.19* | 0.28** | 0.30*** | 0.22** | 0.08** | 0.12** |
| Religiosity | | -0.27 | | 0.43** | | -0.09 | | -0.04 |
| Harshness | | 0.80 | | -1.11 | | -2.31 | | 1.92 |
| $R^2$ | 0.66 | 0.72 | 0.19 | 0.51 | 0.65 | 0.74 | 0.26 | 0.37 |

The dependent variable is indicated in the top of each column. 16 observations. * p-value $\leq$ 10%, ** p-value $\leq$ 5%, *** p-value $\leq$ 1%.