MEMORANDUM

No 12/2014

The Distribution of Individual Conformity under Social Pressure

Moti Michaeli and Daniel Spiro

ISSN: 0809-8786

Department of Economics
University of Oslo

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

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

In co-operation with
The Frisch Centre for Economic Research

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

Last 10 Memoranda

No 11/14	Ingvild Almås, Åshild Auglænd Johnsen and Andreas Kotsadam *Powerty in China as Seen from Outer Space*
No 10/14	Nico Keilman and Coen van Duin *Stochastic Household Forecast by Coherent Random Shares Prediction*
No 09/14	Mads Greaker, Michael Hoel and Knut Einar Rosendahl *Does a Renewable Fuel Standard for Biofuels Reduce Climate Costs?*
No 08/14	Karine Nyborg *Project Evaluation with Democratic Decision-making: What Does Cost-benefit Analysis Really Measure?*
No 07/14	Florian Diekert, Kristen Lund and Tore Schweder *From Open-Access to Individual Quotas:* *Disentangling the Effects of Policy Reform and* *Environmental Changes in the Norwegian Coastal Fishery*
No 06/14	Edwin Leuven, Erik Plug and Marte Rønning *Education and Cancer Risk*
No 05/14	Edwin Leuven, Erik Plug and Marte Rønning *The Relative Contribution of Genetic and Environmental Factors to Cancer Risk and Cancer Mortality in Norway*
No 04/14	Tone Ognedal *Morale in the Market*
No 03/14	Paolo Giovanni Piacquadio *Intergenerational Egalitarianism*
No 02/14	Martin Flatø and Andreas Kotsadam *Drought and Gender Bias in Infant Mortality in Sub-Saharan Africa*

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

The Distribution of Individual Conformity under Social Pressure across Societies*

Moti Michaeli† & Daniel Spiro‡
December 2013

Memo 12/2014-v1
(This version December 2013)

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 an endogenous social norm. The model sheds light on how various punishment regimes induce conformity or law obedience, and by whom, and on phenomena such as societal polarization, unimodal concentration and alienation. In orthodox societies, individuals will tend to either fully conform or totally ignore the social norm, while individuals in liberal societies will tend to compromise between these two extremes. Furthermore, the degree of orthodoxy determines whether those who fairly agree with the norm or those who strongly disapprove it will conform. Likewise, the degree of liberalism determines which individuals will compromise the most. In addition, orthodox societies may adapt norms that are skewed with respect to the private opinions in society, while liberal societies will not do so.

Keywords: Social pressure, Conformity, Liberal, Orthodox, Compliance.

JEL: D01, D30, D7, K42, Z1, Z12, Z13

*We wish to thank Florian Biermann, Sergiu Hart, John Hassler, Arie Kacowicz, Assar Lindbeck, Sten Nyberg, Karine Nyborg, Ignacio Palacios-Huerta, Torsten Persson, Jörgen Weibull, Robert Östling and seminar participants at Hebrew University, University of Oslo, Stockholm University and Stockholm School of Economics for valuable comments.

†Corresponding author. Department of Economics and The Center for the Study of Rationality, Hebrew University, motimich@gmail.com. Moti received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [249159].

‡Dept 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. An early experiment showing the potency of social pressure was done by Asch (1955). He showed that even for seemingly objective issues, such as comparing two lines and stating which one is the longest, social pressure can have strong effects. In economics, models of social norms have been applied to a variety of issues, such as choices of neighborhood (Schelling, 1971), herd behavior (Granowetter, 1978), unemployment (Lindbeck, Nyberg & Weibull, 2003), fertility choices (Manski & Mayshar, 2003) and status seeking (Clark & Oswald, 1998).[1]

This paper studies what is possibly the most basic trade-off that individuals face with regard to social norms and analyzes the aggregate societal outcomes of individuals' decisions across cultures in a heterogenous agent framework.[2] To help fix ideas, imagine a social or political issue that is under some controversy, where there exists a social norm, i.e., a consensual ("right") opinion or norm of behavior. Suppose now that each individual in society has some private opinion regarding this issue, and everyone needs to declare their stance in public. An individual whose private opinion differs from the social norm will then need to consider the trade-off between the social pressure for violating the norm and the psychological cost of stating an opinion different than her private one. In many cases – such as at what age to bear children, how much alcohol to drink and to what extent to follow religious customs – this decision is not binary. Hence, the individual can choose the extent of conformity to the norm from a continuum.

Our mode of analysis, and main aim of the paper, is to examine the *extent* of conformity that one person will exhibit compared to that ex-

[1] Other applications include how much effort to exert under peer pressure (Jones, 1984; and Kandel & Lazear, 1992); and signaling of conformity (Bernheim, 1994). For other topics related to social norms see Goffman (1959) for early research in sociology, Kuran (1995) for political revolutions, Holbrook et al (2003) for effects on political survey making and Hamlin and Jennings (2011) for a review of expressive voting.

[2] Hence, the novelty in our paper comes not in the setup of the basic individual trade-off but from the combination of heterogenous agents who make continuous decisions and from differentiating the results across societies (functional forms). Previous theoretical research with a similar trade-off has been presented by e.g. Brock & Durlauf (2001), Lindbeck et al (2003), Lopéz-Pintado & Watts (2008), Akerlof (1980) and Kuran (1995) who use binary decisions; Bernheim (1994) who has a signaling model with an exogenous norm; Akerlof (1997) who uses peer pressure between three individuals; and Kuran & Sandholm (2008) and Manski & Mayshar (2003) who use peer pressure between many individuals under a double quadratic function.

hibited by another person with a different private opinion. Such comparative statics along the dimension of private opinions provide predictions for (i) which individuals in society will conform more and make larger concessions, (ii) how the distribution of stated opinions in society will look like and (iii) which norms will be sustainable under various societal traits. We show that, although the problem faced by each individual is fairly simple, the outcomes at the aggregate level are quite diverse, and we analyze how they depend on the underlying characteristics of society. The analysis has policy relevance from a normative perspective as it provides predictions on how different punishment regimes (be it social or legal) will affect the extent of compliance across individuals with heterogenous tastes. For instance, it predicts how fines may affect the extent of tax avoidance or illegal driving across individuals, how policies to increase or decrease birthrates may affect individuals with different tastes or how a ban on religious symbols may affect individuals with moderate compared to extreme religious views. From a positive perspective it yields predictions on which societies may sustain skewed norms and whether the public debate will be manifested by polarization, concentration or alienation.

When modelling the characteristics of societies in terms of the social pressure to conform to the norm, one has to take into account that societies in practice differ not only in the general weight of social pressure, but also in its curvature. That is, they differ in the pressure on small deviations from the norm compared to the pressure on large deviations from it. We show that this curvature of the social pressure has more intricate and possibly more important effects than the general weight of pressure. Moreover, in order to connect the model results to outcomes across societies, and based on empirical and casual observations of punishments in groups and societies (to be presented in the next section), we apply labels to the curvature of social pressure: Orthodox societies are those "true to the book" and hence utilize concave social pressure; and Liberal societies are those allowing freedom of expression, as long as it is not too extreme, and hence utilize convex social pressure. Strictly speaking, these labels are not necessary for the formal analysis, but they prove useful and intuitively consistent with actual societies when linking the basic societal characteristics with aggregate outcomes.

The convexity of the social pressure in (what we label as) liberal societies naturally induces individuals to compromise between fully conforming and stating their public opinions. However, depending on the degree of liberalism (i.e., the degree of convexity), the distribution of declared stances will be either bimodal polarization (following the terminology of Esteban & Ray, 1994) or unimodal concentration. Meanwhile, the

concavity of the social pressure in (what we label as) orthodox societies discourages compromise. That is, it will tend to induce individuals to either completely conform or completely speak their minds. Depending on the degree of orthodoxy (i.e., the degree of concavity), we will either see alienation, where those privately opposing the norm partly or totally ignore it, or inversion of opinions, i.e., a case where the norm is maintained by those opposing it the most. Hence, while it may seem intuitive that people whose private opinions are close to the norm should display more conformity than those whose private opinions are further away, we show that this is not always true. In fact, for a wide set of preferences (in both liberal and orthodox societies) the opposite holds – opinions are inverted, so that those who dislike the norm the most adhere to it more than others.

Another outcome that clearly separates orthodox and liberal societies is that when the norm is endogenized to represent the average *declared* opinion in society, in liberal societies it will also represent the average *private* opinion.[3] In contrast, in orthodox societies we may well obtain a skewed social norm centered on a point that is far from what people really think. This provides predictions for when we should (and when we should not) expect to find norms that are unrepresentative of the private preferences.[4]

An overarching analytical result is that the curvature of social pressure relative to that of inner preferences determines who in society is most affected by social pressure. More precisely, if the concavity of social pressure (arising from deviations from the norm) is higher than that of the cognitive dissonance (from deviating from one's bliss point), then individuals with inner preferences close to the social norm will concede (i.e. move towards the norm) relatively more than those with private preferences far from it, and vice versa. Roughly speaking, this means that the more orthodox a society is, the more directed it will be at making individuals who privately almost agree with the norm completely conform to it, while alienating others. Likewise, the more liberal a society is, the more directed it will be at making individuals who strongly disagree with the norm conform at least a little bit, while hardly affecting the stances of those who privately tend to agree with the norm.

The next section describes some observations of punishments across

[3] This is true at least, but not only, if the distribution of true opinions is uniform.

[4] This can be contrasted to the well known results of Brock & Durlauf (2001) who study binary decisions. They find multiplicity of equilibria under nearly all circumstances – only the weight of social pressure matters. We find that when allowing for a continuous decision variable (e.g. how much to cheat on taxes, how early to bear children or how fast to drive on the motorway), the multiplicity of equilibria and norms appears only in orthodox societies.

societies and suggests some labels distinguishing them. Then the model and analytical definitions are outlined in section 3. Sections 4 and 5 are devoted to analyzing (what we label as) liberal and orthodox societies respectively. They analyze how individual conformity varies as a function of private opinions, what the distribution of stances will be in society as a whole, and what the endogenous location of the norm will be. Section 6 presents the overarching result on relative conformity. Finally section 7 concludes and discusses the outcome differences between liberal and orthodox societies. To keep the paper readable, the more elaborate analytical derivations and proofs are covered in the appendix.

2 Social pressure across cultures

In a recent empirical paper Gelfand et al. (2011) construct a measure of the "tightness" of societies that is described as the "overall strength of social norms and tolerance of deviance". For instance, they find that India, Pakistan and Malaysia have strong norm enforcement while Hungary, Estonia and Ukraine have the weakest enforcement. Such a measure of norms clearly provides an important distinction between societies. But while measuring norm enforcement using a single measure may be sufficient in some instances, in other settings it will be important to also distinguish between how a society sanctions small deviations in comparison to larger ones. While one society may impose harder social pressure for small deviations from the norm compared to a second society, the second may punish large deviations harder than the first one – they differ in the curvature of punishment. As we will show in this paper, this distinction is important in terms of predicting aggregate outcomes.

In order to show that societies do differ along this dimension also in practice, a few observations may be in place. An example comes from experiments using the Public Good Game that have been performed by Herrmann et al. (2008). They document how participants sanction others who contribute a different amount than themselves to a public good. By ocular inspection (see figure 1 in Herrmann et al. 2008), their results suggest that deviations are punished convexly in places such as Copenhagen, Bonn and Melbourne while being punished concavely in places such as Muscat and Riyadh. Another detail to note in their results is that Melbourne has heavier punishments for large deviations than has either Riyadh or Muscat, while it has lighter punishments than these two for small deviations. This pattern matches that of the stylized societies 2 (representing Muscat and Riyadh) and 3 (representing Melbourne) in Figure 1. Another recent experiment, by Krupka & Weber (2013) reveals

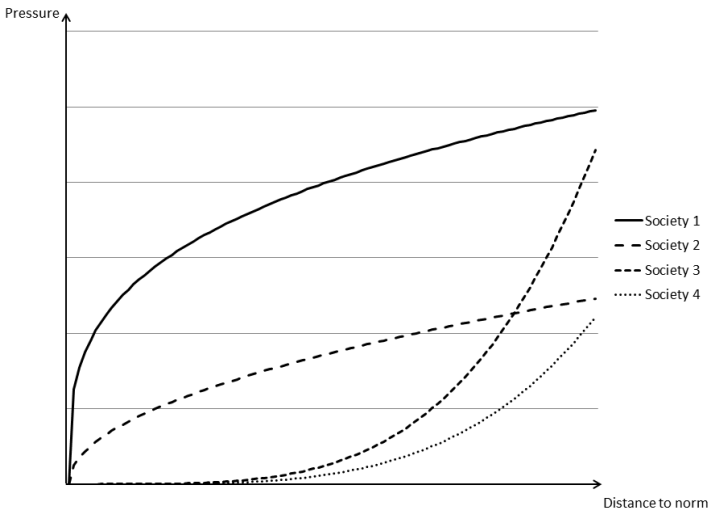

Figure 1: Punishment across societies. A system of punishments may be at the same time harsh and concave (society 1). Alternatively, it may be harsh and convex (society 2). Or, it may be light and concave (society 3). Finally, like in society 4, it may also be light and convex.

a concave social pressure in the dictator and bully games.[5]

A more anecdotal demonstration of these points is by crudely comparing the punishment system in the Israeli Jewish Ultraorthodox community, with the punishment system under the Taliban regime on the one hand, and with liberal West European institutions on the other hand. This is to some extent a comparison of informal and formal sanctioning, but the purpose here is to highlight that punishment systems can distinguish between large and small wrongdoings in different ways.[6]

An important difference between the Taliban and the Ultraorthodox sanctioning systems is that the Taliban have substantially heavier punishment for any comparable deviation from the norm. There are numerous accounts of the Taliban using capital punishment for both misdemeanor and larger offenses, while the Ultraorthodox are characterized by milder punishments, such as censuring or at most excommunicating norm violators. However, one characteristic that both these societies

[5] See Krupka & Weber's (2013) Figures 3 and 5. Applied to the discussion here their graphs should be inverted since their measure is of "appropriateness of behavior" which then is the inverse of our vocabulary which is about "inappropriateness".

[6] Hence, while there are many dimensions by which these societies can be compared, for the purpose of this paper we will focus only on two aspects – the curvature of punishment and its general weight.

have in common is that they advocate "being true to the book", meaning that they will sanction relatively harshly (compared to their own scale) any small deviation from the norm, while large deviations will be sanctioned only slightly more. For instance, in the Israeli Ultraorthodox society, a woman may be censured for wearing a dress that is too short, and a man for publicly supporting the draft of Ultraorthodox to the Israeli army.[7] Furthermore, there is an upper bound on the severity of punishment. In the Ultraorthodox society this follows from the fact that individuals are no longer punishable after being excommunicated from the community, and in the Taliban it follows from the terminal nature of capital punishment which they seem to apply rather generously. The difference between the Taliban and the Jewish Ultraorthodox when it comes to punishment is like the difference between societies 1 and 2 in Figure 1 – both are concave but one is harsher at all levels.

What about the punishment structure of the liberal West European institutions? Virtually by definition, a liberal democracy must allow the expression of (almost) any view. By most democratic constitutions, citizens are allowed a rather broadly encompassing freedom of expression and are eligible to run for elections using almost any political platform. But once a party or an individual expresses views that deviate very far from the consensus, for example a party that wants to abolish democracy or an individual that incites people to commit crimes, that party may become illegal (like Nazi parties are in certain countries), and that individual may be fined or arrested (for incitement) or be subject to surveillance (if she, for example, openly expresses extreme right-wing or extreme left-wing opinions, or supports Sharia Law). Roughly speaking, this means liberal democracies will tend to be convex in how they constitutionally deal with deviations, like societies 3 and 4 in Figure 1.

As incomplete and stylized as these descriptions may be, they do highlight that representing a punishment system with only one parameter is insufficient. There is on the one hand the issue of harshness in general, and there is, on the other hand, the issue of curvature of the punishment system as can be captured by varying the degree of convexity or concavity.

Can we connect these mathematical terms to more common descriptions relating to societies and cultures? Connecting precise mathematical definitions into labels useful in social and behavioral sciences always comes with certain subjectivity. But we find labels useful in interpreting the mathematical results and providing intuition. For a formal analysis

[7]It should be emphasized here that these examples are minor misdemeanors in terms of the Jewish religion as religious Jews belonging to less extreme factions treat these behaviors as completely normative.

such labeling is naturally redundant, as the mathematical results will hold whatever labels we use. It is furthermore conceivable that a society may employ different punishment structures for different issues. Hence our labels should be interpreted as describing one dimension of one society or group.

To represent the general strength of the norm we adopt the term "Tightness" from the vocabulary of Gelfand et al. (2011). This means that, holding other things equal, a tighter society will have a harsher social pressure. As for the curvature of the social pressure, we believe that a concave pressure – being meticulous about minor deviations from the norm but not distinguishing so much between large and small wrongdoings (i.e., advocating behavior that is "true to the book") – represents "Orthodox" societies. This also seems to be in line with the experiments performed by Hermann et al (2008) in Riyadh and Muscat, which we would intuitively label as orthodox. Likewise, it seems reasonable to characterize both the Israeli Ultraorthodox community and the Taliban regime as orthodox.

As a counter label, following our descriptions of how liberal democracies constitutionally treat the freedom of expression, and in order to portray the examples of Melbourne, Bonn and Copenhagen in Herrmann et al. (2008), we suggest the label "Liberal" to represent societies with convex social pressure. That is, in these societies small deviations are largely ignored, but for sufficiently large deviations the pressure is ramped up.

It should be noted that in everyday language (and perhaps even in the scientific one), there is an overlap of the terms "orthodox" and "tight". Both are often taken to imply the existence of rather strong social pressure. This blend of terms reflects that orthodox societies possibly use heavier social pressure compared to liberal ones.[8] However, while in practice there may be a positive correlation between tightness and orthodoxy, this should not necessarily always be the case, as demonstrated earlier by comparing the punishments imposed for large deviations by subjects in Melbourne and in Muscat in the experiments of Herrmann et al. (2008).

Just as societies may be characterized by the gradual change in punishments in treating norm deviations, individuals may be characterized by different sensitivities to the psychological cost embodied in small misrepresentations as compared to larger ones when stating stances in public. The discomfort of the individual may then rise concavely or convexly

[8]There is also a more "technical" reason for this association of terms. If one compares the pressures of societies 1 and 3 in Figure 1, one may notice that although the graphs start and end almost at the same points, the concave graph is clearly high above the convex one, as follows from the basic characteristics of these functions.

the further the expressed opinion is from her private one. Theoretically we see no particular reason why a convex psychological cost function would be more or less reasonable than a concave. While in previous theoretical research a convex disutility is more common (e.g. Bernheim, 1994; Manski & Mayshar, 2003; Clark & Oswald, 1998) some recent experimental research suggests concave preferences may be present in many cases too (e.g. Gino et al 2010; Gneezy et al, 2013; Kendall et al, 2013).

We will use the term "Perfectionist" to describe individual punctiliousness. That is, perfectionist individuals will be those who are very reluctant to state opinions or perform actions that deviate even slightly from their ideology, but once they do deviate slightly from their ideological bliss point, any further deviations make little difference. We will use the counter-label "Non-perfectionist" to represent the individual trait of a convex displeasure of deviating from the private opinion. Then, as long as the deviation is not too large, it hardly inflicts any discomfort.

3 The model

An individual is represented by a type $t \in (t_l, t_h)$, 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

$$D(t-s), \ \frac{dD}{d(|t-s|)} > 0,$$

If a person minimizes D only, it is immediate that $s(t) = t$. This way t represents the bliss point of an individual in fulfilling her private preferences and D can be interpreted as the cognitive dissonance or displeasure felt by taking a stance that is not in line with this bliss point. We can, for example, think of t as the position on an ideological scale. $s(t)$ can then be interpreted as an ideological statement or an ideological action such as the extent of adherence to a religious custom.

However, an individual who takes s as a stance, also feels a social pressure $P(s-\bar{s})$, where $\bar{s}$ can be understood as a social norm which is exogenous from the point of view of an individual. In equilibrium it will be determined by the average stance in society. From the individual's point of view we have

$$\frac{dP}{d(|s-\bar{s}|)} > 0.$$

The total disutility (or loss) of an individual is the sum of the cognitive

dissonance and the social pressure,

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

Seeking to minimize $L(t,s)$, it is immediate that each individual will take a stance somewhere (weakly) in between her private bliss point and the social norm. 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.

To compare the extent of norm conformity for different individuals in society, we will use two different measures.

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 impose negativity so that conformity is increasing the closer s^* is to $\bar{s}$. That is, for $t \geq \bar{s}$ conformity is locally weakly decreasing in t if and only if $\frac{ds^*}{dt} \geq 0$. 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 how much an individual is giving up on her private opinion compared to how much she could, maximally, 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 main analysis revolves around the function $s^*(t)$. Nearly all upcoming results can be derived by using general convex and concave functional forms, but for brevity and in order to facilitate the interpretation, we will assume that the cognitive dissonance and social pressure are power functions.

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

These functions are symmetric around $t = s$ and $s = \bar{s}$, respectively. For conservation of space, we will therefore mainly only present the problem and solution for $t \geq \bar{s}$, where we get the following minimization problem:

$$\min_s \left\{ (t-s)^{\alpha} + K(s-\bar{s})^{\beta} \right\},$$

with a first-order condition

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

and a second-order condition for an internal local minimum point

$$(\alpha-1)\,\alpha (t-s)^{\alpha-2} + (\beta-1)\,\beta K (s-\bar{s})^{\beta-2} > 0. \tag{3}$$

Following the previous section we will use the label Liberal for $\beta > 1$ and Orthodox for $\beta < 1$. The limiting case $\beta = 1$ can be seen as weakly orthodox and weakly liberal at the same time. Likewise we call individuals Perfectionist when $\alpha < 1$ and Non-perfectionist when $\alpha > 1$ (when $\alpha = 1$ the individual is perfectionism-neutral). Finally, K represents the Tightness of society in relative terms, i.e. , relative to the size of the cognitive dissonance, which is normalized to 1 (we will also refer to K as the *weight* of social pressure). We will assume that the only difference between individuals is in their bliss points while having α, β and K in common.

When presenting results about the distribution of stances in a society we also need to specify a distribution of types. To make this as transparent and neutral as possible we will present the stance distribution when $t \sim U(t_l, t_h)$.

In total, this provides a rich description of societies (or cultures) in terms of their basic characteristics and outcomes. Each society has its underlying characteristics made of the distribution of private opinions, the curvature of social pressure, the curvature of the psychological cost and the weight of pressure. In each society, one 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 norms a society can sustain.[9]

4 Liberal societies

We start by examining the case when β is greater than 1. From the second-order condition (3), it is immediate that there is an internal solution for every type t if $\alpha \geq 1$ and a possibility for both inner and corner solutions when $\alpha < 1$. The properties of stances and conformity in society are summarized in the following proposition.[10]

[9] Kuran & Sandholm (2008) define a culture as the distribution of private opinions and the distribution of stated opinions. However, restricting themselves to double quadratic functions they do not let the curvature vary by society. As we show in the paper, the distinction by curvature is an important driver of societal outcomes and behavior.

[10] We ignore the special case of $\alpha = \beta$ as it is a borderline case between $\alpha < \beta$ and $\alpha > \beta$, where $|s^*(t) - \bar{s}|$ is linearly increasing and the distribution of s^* is uniform.

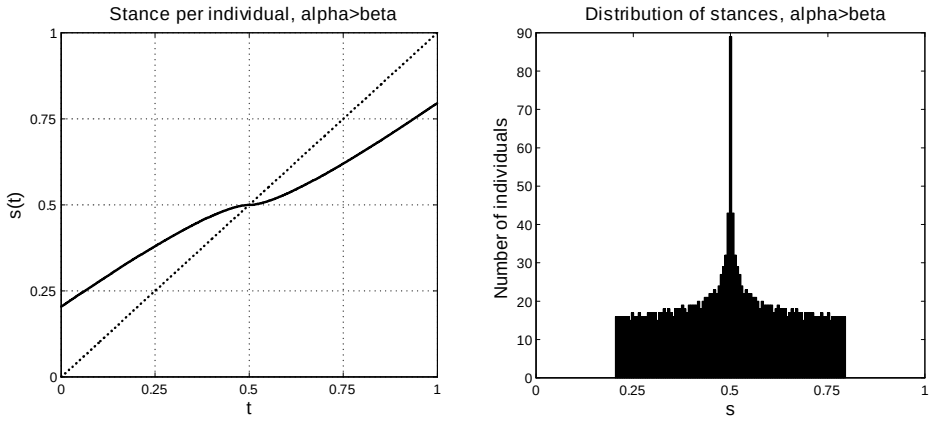

Figure 2: $1 < \beta < \alpha$ with $\bar{s} = .5$, $t \sim U(0,1)$. The left-hand schedule depicts $s^*(t)$ (full line) and compares it to the line $s = t$ (dashed line). The right-hand schedule depicts the probability density function.

Proposition 1 *If $\beta \geq 1$ and $\bar{s} \in]t_l, t_h[$ then:*

1. *If $\alpha > \beta$, then $|s^*(t) - \bar{s}|$ is convexly increasing in $|t - \bar{s}|$, thus conformity is decreasing in $|t - \bar{s}|$. Moreover, the relative concession is decreasing in $|t - \bar{s}|$. If, furthermore, $t \sim U(t_l, t_h)$, then the distribution of s^* is unimodal.*

2. *If $1 \leq \alpha < \beta$, then $|s^*(t) - \bar{s}|$ is concavely increasing in $|t - \bar{s}|$, thus conformity is decreasing in $|t - \bar{s}|$. Moreover, the relative concession is increasing in $|t - \bar{s}|$. If, furthermore, $t \sim U(t_l, t_h)$, then the distribution of s^* is bimodal.*[11]

3. *If $\alpha < 1$ and the range of types is broad enough, then $|s^*(t) - \bar{s}|$ is first increasing then decreasing in $|t - \bar{s}|$. The relative concession is increasing in $|t - \bar{s}|$. If, furthermore, $t \sim U(t_l, t_h)$, then the distribution of s^* is bimodal.*

Proof. *See appendix.* ■

The results are visualized in Figures 2, 3 and 4 where the left-hand schedules represent the resulting function $s^*(t)$ and the right-hand schedules represent the resultant distribution (the probability density function) given a uniform distribution of bliss points.

When $\beta > 1$ only extremists (t far from $\bar{s}$) feel any substantial social pressure to deviate from their bliss point. Then, if $\alpha > 1$, an individual's

[11] If $\alpha = 1$ the final statement is contingent on a sufficiently large K and a sufficiently centered norm.

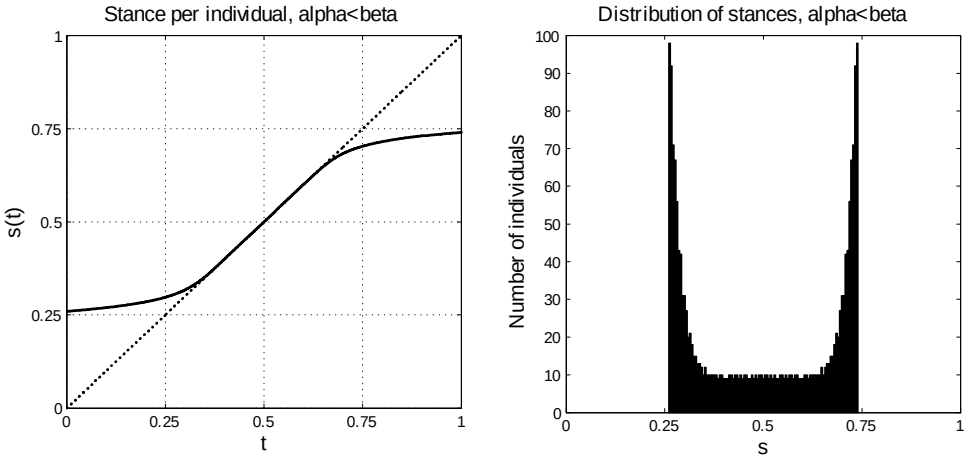

Figure 3: $1 < \alpha < \beta$ with $\bar{s} = .5$, $t \sim U(0,1)$. The left-hand schedule depicts $s^*(t)$ (full line) and compares it to the line $s = t$ (dashed line). The right-hand schedule depicts the probability density function.

inner preferences are also open for deviations from her bliss point, as long as the deviation is not too large. The important question is then whether it is the cognitive dissonance or the social pressure that is more open to deviations.

To get the intuition for the first part of the proposition ($\alpha > \beta$), it may be easiest to imagine a very large α. Then, an individual does hardly feel any dissonance from deviating a little from t. A moderate person (i.e., t close to $\bar{s}$) may then just as well choose a stance very close to $\bar{s}$ in order to minimize the social pressure. However, an extreme type (i.e., t far from $\bar{s}$) will not be willing to move equally close to $\bar{s}$ since the inner discomfort will then be very large. Thus, in this scenario, moderates tend to concede relatively more to the norm. The resultant distribution will therefore be a concentration of individual statements around the norm.

When $\beta > \alpha$ (the second result in the proposition), it may be easiest to imagine a very large β. Then, an individual does hardly feel any pressure by deviating a little from $\bar{s}$. Consequentially, only extreme types will feel enough social pressure to actually take a large step from their bliss point. Meanwhile moderates will hardly be inclined to move from their bliss points. There will then be a concentration of extreme types at a certain distance on each side of the norm. That is, society will be polarized.[12]

[12]Strictly speaking, this is true if and only if the norm is sufficiently centrally located, which we will show to be the case in equilibrium.

The polarization represents the result that, in very liberal societies, extremists make relatively large concessions. Hence, while all liberal societies (consisting of non-perfectionists) have in common that they get all individuals to compromise, the degree of liberalism determines who will concede more. In less liberal societies ($\beta < \alpha$) moderates will make larger concessions creating unimodal concentration, while in very liberal societies ($\beta > \alpha$) the extremists will concede more. This latter case creates polarization in the sense of Esteban & Ray (1994), where there are concentrations of public statements at two different locations on the axis. If we fix β and let α fall gradually, the population becomes more polarized with a higher concentration at the peaks and less individuals taking intermediate stances (the "smile" in Figure 3 becomes deeper).

The continuation of this logic of gradually decreasing α is represented by the third statement in the proposition. As individuals pass the threshold from non-perfectionist ($\alpha > 1$) to perfectionist ($\alpha < 1$) we get an enhancement of the previous logic. In general, individuals with $\alpha < 1$ will either not concede at all, or, once they cannot declare their private opinion, concede a great deal. In a liberal society social pressure is hardly present for small deviations implying moderates will declare exactly their type. But since social pressure is ramped up for large deviations it will be hard for extremists to stand firmly at their bliss point. Hence, being perfectionists, they may as well conform a lot. This means that some types (moderates) will not make any concession while others (extremists) will make very large concessions. As is illustrated in Figure 4 (left-hand schedule), the proposition then implies that the extremists conform more than some moderates and that, within the group of those conforming, the more extreme individuals are the most conformed. Thus, we get what can be called an inversion of opinions, where extreme people conform more than moderates. Furthermore, we get it at two levels. Between extremists and moderates the extremists conform more and within the group of extremists the most extreme conform more than the less extreme. All in all, this will create a bimodal distribution (Figure 4, right-hand schedule) where extremists form the peaks and there is a uniform distribution of moderates around the peaks – society will seem polarized. As α increases towards 1, these peaks will move outwards and as α passes 1, the inversion ceases to exist and we are left with the same bimodally distributed society as in the case of $1 \leq \alpha < \beta$ (see Figure 3).

Up until now we have described the results for any norm, no matter where it is located. But what would be possible equilibrium locations of a social norm in a liberal society? For this we need to specify how the norm location is determined in society. The norm being the average

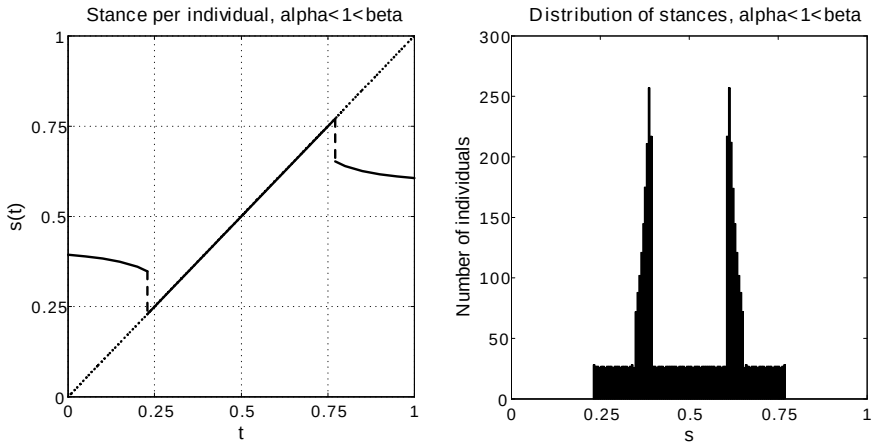

Figure 4: $\alpha < 1 \leq \beta$ with $\bar{s} = .5$, $t \sim U(0,1)$. The left-hand schedule depicts $s^*(t)$ (full line) and, in comparison, the line $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function.

stated opinion seems like a relevant possibility. That is

$$\bar{s} = \frac{1}{t_h - t_l} \int_{t_l}^{t_h} s^*(\tau)\, d\tau.$$

Naturally, there may be other forces shaping the norm, but the average stance seems like a natural way to start, which has also been applied in earlier research (e.g. Clark & Oswald, 1998; Brock & Durlauf, 2001).[13]

In the following analysis, we implicitly assume that the distribution of *types* is uniform. This implies a distribution of stances (S) which is (at least locally) symmetric around $\bar{s}$. Thus, for a certain social norm to be the average of *all* declared stances, S has to be symmetric around $\bar{s}$ over the whole range of types. This condition holds immediately for a centered norm. I.e., a norm at the average bliss point ($\frac{t_h+t_l}{2}$) is sustainable in equilibrium. But it does not hold for a skewed norm unless types with opinions far from it fully conform (if they do not fully conform, they pull the norm in their direction). Since in a liberal society no one fully conforms, the following proposition holds.

Proposition 2 *Suppose $\bar{s}$ is the average stance in society and $t \sim U(t_l, t_h)$. Then, if $\beta > 1$, there is a unique equilibrium where $\bar{s} = \frac{t_h+t_l}{2}$.*

[13] In two adjacent papers we investigate how pressure and norms are formed in societies where either the stated or the private opinions of individuals determine social pressure (see Michaeli & Spiro, 2013a,b).

The proposition states that the only possible equilibrium in a (strictly) liberal society is where the social norm is equal to the average bliss point. Thus, the norm is bound to be representative of the actual private opinions in society.

5 Orthodox societies

When $\beta \leq 1$, society is intolerant to small deviations from the consensus, but hardly distinguishes between moderate and large deviations. It is now immediate from the second-order condition (3) that if $\alpha \leq 1$, then any inner solution is a maximum, implying that individuals will either fully conform ($s^*(t) = \bar{s}$) or speak their minds ($s^*(t) = t$). This is also intuitive, because when the functions are concave, taking a stance in between t and $\bar{s}$ would imply both great dissonance and heavy social pressure. The heart of the matter is who chooses to fully conform and who chooses to speak her mind in public. It turns out that there exists a distance from the norm, $\Delta \equiv K^{\frac{1}{\alpha-\beta}}$, at which the optimal solution switches between these two corner solutions. Thus, if the range of types is broad enough, conformity changes drastically at $\bar{s} \pm \Delta$. When $\alpha > 1$, types close the norm have corners solution too, but types further enough from it choose a compromise solution.[14] The properties of stances and conformity in society for this case are summarized in the following proposition.[15]

Proposition 3 *Let $\Delta \equiv K^{\frac{1}{\alpha-\beta}}$. If $\beta \leq 1$ and the range of types is broad enough, 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 $|t - \bar{s}|$. If, furthermore, $t \sim U(t_l, t_h)$ and $\bar{s}$ is sufficiently centered, then the distribution of s^* is unimodal and discontinuous with 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*

[14] It can be shown that the distance from the norm to the switching points from corner solutions to inner solutions in this case is smaller than Δ.

[15] We ignore the special case of $\alpha = \beta$ as it is a borderline case between $\alpha < \beta$ and $\alpha > \beta$, where all types either speak their minds (if $K < 1$) or fully conform (if $K > 1$), and so the distribution of s^* is either uniform (if $K < 1$) or degenerate at $\bar{s}$ (if $K > 1$).

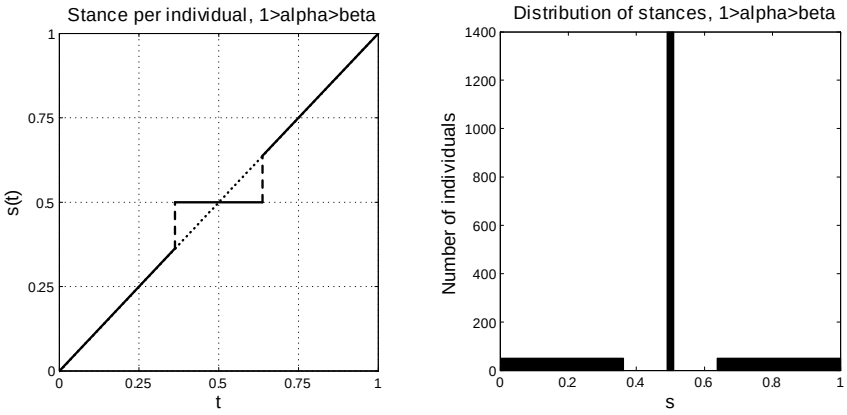

Figure 5: $\beta < \alpha \leq 1$ with $\bar{s} = .5$, $t \sim U(0,1)$. The left-hand schedule depicts $s^*(t)$ (full line) and in comparison, the $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function.

increasing in $|t - \bar{s}|$. If, furthermore, $t \sim U(t_l, t_h)$, then the distribution of s^* is unimodal and continuous with a peak at $\bar{s}$ and a uniform section attached to it.

3. *If $\alpha > 1$, then types close enough to the norm fully conform, thus $|s^*(t) - \bar{s}| = 0$ at that range. For types far from the norm $|s^*(t) - \bar{s}|$ is increasing in $|t - \bar{s}|$. Conformity and relative concession are weakly decreasing in $|t - \bar{s}|$. If furthermore, $t \sim U(t_l, t_h)$ and $\bar{s}$ is sufficiently centered, then the distribution is discontinuously trimodal with a central peak at $\bar{s}$ and a detached group on each side.*

Proof. *See appendix.* ■

In part 1 of the above proposition, society displays a relatively high degree of orthodoxy ($\beta < \alpha$). Individuals with opinions close enough to the social norm ($t \in [\bar{s} - \Delta, \bar{s} + \Delta]$) will choose to fully conform while individuals with opinions far enough from the norm will simply cope with the full social pressure and choose the inner bliss point as their stance. The intuition is that in (very) orthodox societies one has to move all the way to the norm to alleviate pressure to any substantial degree. Then, when $\alpha > \beta$, extreme types (t far from $\bar{s}$) find it relatively more painful to move to the norm compared to stating their type. Altogether, this means that very orthodox societies create an either-or mentality which alienates people with opinions far from the norm, but compels those with opinions close enough to it to fully align. If a person feels that it is not possible to fulfill the norm, there is no point in trying to partly conform,

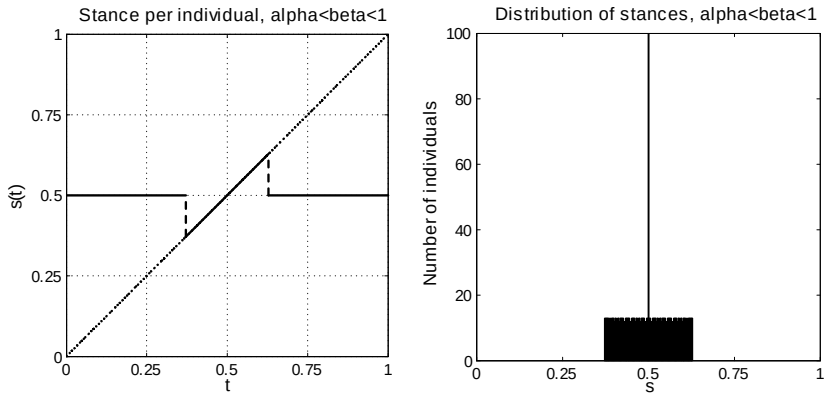

Figure 6: $\alpha < \beta \leq 1$ with $\bar{s} = .5$, $t \sim U(0,1)$. The left-hand schedule depicts $s^*(t)$ (full line) and compares it to the line $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function.

since this will hardly make a difference anyway. This way an orthodox society, which is not tolerant to small deviations from the norm, will tend to fail in moderating extreme people's stances.

We will now continue with an inversion of the previous case, which holds when society is orthodox to a lesser degree, so that $\beta > \alpha$ (part 2 of Proposition 3). The observable outcome of this case is a distribution that looks like a standard concentration of individuals at the norm. But there is an important twist. The concentration of stances at $\bar{s}$ consists of individuals with extreme inner bliss points, i.e., those with private opinions far from the norm. Thus, the extreme types' declarations are more conformed than those of the moderates. This means that as the distance from the norm increases, conformity is initially decreasing, but then sharply increases at the switching point from totally ignoring the norm to full conformity. This creates a form of inversion of opinions where those who despise the norm the most are the (only) ones upholding it. The intuition is that moderates are now unwilling to conform since this would inflict too great displeasure given that the dissonance is so concave. For extremists, however, not conforming will imply too great social pressure since $P(t, \bar{s})$ is increasing relative to $D(t, \bar{s})$ with the distance from the norm ($|t - \bar{s}|$). This means that mildly orthodox societies will be good at attracting extremists to the norm while "allowing" the freedom of expression of those (sufficiently) close to it. 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.

By comparing part 1 and part 2 of the proposition, we see that both types of orthodox societies with perfectionist individuals have one thing

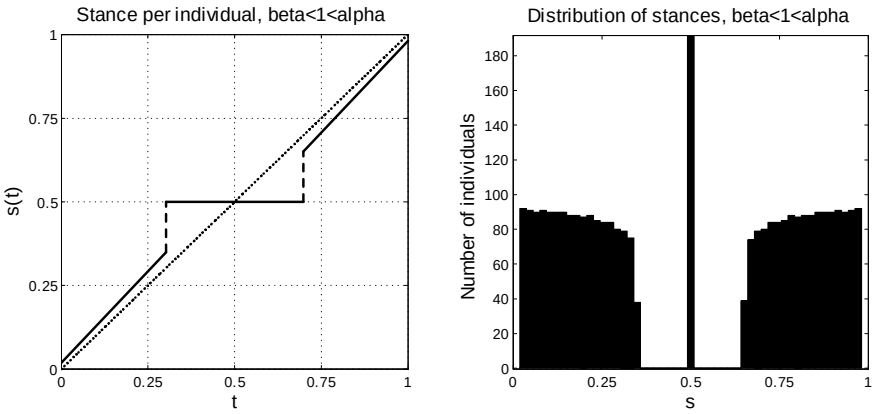

Figure 7: $\beta < 1 \leq \alpha$ with $\bar{s} = .5$, $t \sim U(0,1)$. The left-hand schedule depicts $s^*(t)$ (full line) and, in comparison, the line $s = t$ (dashed line). The right-hand schedule depicts the probability distribution function. Note that the y-axis is truncated from above for visual purposes.

in common – making each person either conform fully or not at all. But the degree of orthodoxy (i.e., whether $\beta > \alpha$ or $\beta < \alpha$) makes a refinement of this result by yielding completely opposite predictions when it comes to which individuals will be the ones conforming (extremists or moderates respectively) and what the distribution of stances will be (unimodal concentration or detachment at the extremes respectively).

In the third part of the proposition, when $\alpha > 1$, individuals are non-perfectionist, so only large deviations from the bliss point create dissonance. 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.[16] Figure 7 illustrates the resultant distribution of stances. It has a peak at $\bar{s}$ and (for a sufficiently broad range of types and a sufficiently centered $\bar{s}$) has two detached groups towards the extreme ends.[17]

[16] It is generally hard to find a closed form solution for the cutoff between conformity and inner solutions in this case. However, the inner solution is increasing relative to the corner solution as t is distanced from the norm. Hence, for a broad enough range of types, we know that the inner solution is preferred by extreme individuals. In contrast, and perhaps trivially, the cutoff is increasing in K in such a way that if the social pressure has enough weight, there can arise a case of everyone choosing the norm.

[17] The requirement for a sufficiently broad range of types can alternatively be replaced with a requirement for sufficiently small K. I.e., we can get qualitatively similar societies by either adding heterogeneity (broadening the range of types) or by decreasing the weight of punishment (decreasing K).

The intuition for this is that, since society is orthodox (concave social pressure), small deviations from the norm draw relatively heavy pressure. Combining this with non-perfectionism at an individual level – small deviations from the bliss point are almost painless – implies that moderates do best in completely conforming to the social norm. In comparison, extremists would feel too much dissonance if they were to fully conform. However, since the dissonance is convex, the extremists do not mind making small concessions. Hence, they choose a compromise solution. Naturally, the more extreme types will conform even less. As can be seen, this closely resembles the case where society is very orthodox and individuals are perfectionist. Also here extremists are alienated from society, but instead of completely "ignoring" the norm, they make small concessions to adapt to it.[18] The general lesson from these two cases is that orthodoxy creates alienation if individuals are not very perfectionist. One interpretation is that individuals who strongly object the norm in an orthodox society will prefer to be excommunicated, like happens, for instance, with dissenters in the Jewish Ultraorthodox community.

Let us now analyze which social norms can be sustained in equilibrium if the norm is the average stance in society. Recall that when the distribution of types is uniform, the distribution of stances (S) is locally symmetric around $\bar{s}$. Thus, for a certain social norm to constitute an equilibrium, S has to be symmetric around $\bar{s}$ when including the stances of all types in the range $[t_l, t_h]$. As was explained in Section 4, this condition holds immediately for a norm at the center of the types' distribution ($\frac{t_h+t_l}{2}$), but does not hold for a skewed norm unless types with opinions far from it fully conform. Since in orthodox societies we do find scenarios where such types fully conform, those societies may sustain a skewed norm in equilibrium, as stated in the next proposition.[19]

Proposition 4 *Let* $\Delta \equiv K^{\frac{1}{\alpha-\beta}}$. *Suppose* $\bar{s}$ *is the average stance in society and* $t \sim U(t_l, t_h)$. *Then if* $\beta \leq 1$, *there exists an equilibrium where* $\bar{s} = \frac{t_h+t_l}{2}$. *Furthermore:*

1. *If* $1 < \alpha$ *then* $\bar{s} \neq \frac{t_h+t_l}{2}$ *can be sustained in equilibrium if and only if* $s^*(t) = \bar{s}\ \forall t$.

[18] This is similar to the result in Bernheim's (1994) paper. What is interesting is that while Bernheim gets this distribution when both pressure and dissonance are convex functions (according to our way of defining them), we get it when dissonance is convex but pressure is concave. This way, whether pressure is applied to actions (our model) or beliefs about types (Bernheim's model) makes an important difference.

[19] We ignore the special case of $\alpha = \beta$. Then any $\bar{s} \neq \frac{t_h+t_l}{2}$ can be sustained in equilibrium if and only if $K \geq 1$.

2. *If* $\beta < \alpha \leq 1$ *then* $\bar{s} \neq \frac{t_h+t_l}{2}$ *can be sustained in equilibrium if and only if* $t_h - t_l < 2\Delta$ *and* $\bar{s} \in [t_h - \Delta, t_l + \Delta]$.

3. *If* $\alpha < \beta$ *then* $\bar{s} \neq \frac{t_h+t_l}{2}$ *can be sustained in equilibrium if and only if* $t_h - t_l > 2\Delta$ *and* $\bar{s} \in [t_l + \Delta, t_h - \Delta]$.

Proof. *See the appendix.* ■

Parts 1 and 2 of the proposition deal with cases in which (only) moderate individuals fully conform. Therefore, in order for a skewed norm to hold in equilibrium, the range of types needs to be narrow enough to include no "extremists". Otherwise types far from the norm will be alienated, creating opposition on one extreme end of society and hence unbalancing the norm. Another perspective on this result highlights the role of the severity of the social pressure to conform. If the norm is skewed towards one of the edges, a higher pressure weight K (i.e. a tighter society) is needed to make extremists at the far end conform. Hence, a larger K enables more skewed norms. As the weight of pressure falls, this leaves a narrower range of possible equilibria, and eventually, the only remaining equilibrium is when the social norm is equal to the average bliss point. This implies that in very orthodox societies, the only way of upholding a skewed social norm is by either having severe social pressure (i.e., creating cohesion of statements) or by having individuals with a tight range of bliss points (i.e., having cohesive private opinions).

The third part of the proposition deals with the case in which the "extreme" individuals are those who fully conform. As shown earlier, this happens in orthodox societies with very perfectionist individuals. Such a society, albeit being orthodox, in practice creates freedom of expression for those close to the norm, since inner preferences are very concave. However, the potential dissenters, in terms of their private opinions, find the social pressure too strong to resist, and perfectionist as they are, end up fully conforming. By doing so they give up their say in determining its location, thus enabling the existence of norms that are unfavorable to themselves. It may be interesting to note that, unlike the case of $\beta < \alpha$, it is here not necessary to have cohesive private opinions for a norm to be sustained. In fact, on the contrary, a broader range of types enables a broader range of norms, as can be seen in the equilibrium conditions for $\bar{s}$. But an important feature of this equilibrium is that there is equally much mild critique coming from both sides of the norm – the debate in society has to be balanced around the norm. Note also that for any finite K there is always a share of individuals with private opinions close to the norm who speak their mind. Since equally many

of these need to be on each side of the norm, this sets a limit for how skewed the norm may be.

6 Relative concession across societies

By now it should be already clear that it is not only the convexity (or concavity) of the social pressure and of the cognitive dissonance that matter – their *relative curvature* plays a significant role too. In Section 4 we saw that a change in the relative curvature (i.e., which of α and β is the greatest) implies a switch between unimodal and bimodal distributions, and in Section 5 we saw that a change in the relative curvature implies a switch between a state of alienation and a state of inversion of opinions. These results can be generalized to an overarching pattern of who in society (moderates or extremists) concedes relatively more to the social pressure.

Corollary 5 *The relative concession is decreasing in* $|t - \bar{s}|$ *if and only if* $\beta < \alpha$.
Proof. *Follows directly from propositions 1 and 3.* ■

This corollary establishes that when the social pressure is more concave (or less convex) than the cognitive dissonance, it mainly affects moderates.[20] This is intuitive since, roughly speaking, when the pressure is relatively concave, then small deviations from the norm matter more than large deviations, and so small concessions by moderates matter more (in terms of relieving the pressure) than small concessions by extremists. If the converse holds, such that the social pressure is relatively more convex, it mainly induces the extremists to conform, since small concessions far from the norm matter more than small concessions close to it. This result has an implication for the distributions of stances that will be formed in different societies. Within the orthodox societies, those societies that are more orthodox will be more directed at creating conformity among people who privately almost agree with the norm. This will be at the expense of moderating those who strongly disagree with the norm. As for liberal societies, the same result implies that the more liberal among them will be more directed at getting extremists to make compromises, while affecting the stances of moderates to a lesser extent.[21]

[20] With general functional forms, the condition for decreasing relative concession is generalized to $\gamma_P(x) \equiv -\frac{xP''(x)}{P'(x)} > \gamma_D(x) \equiv -\frac{xD''(x)}{D'(x)}$, where $\gamma_F(x)$ is the Arrow-Pratt measure of relative risk aversion of the function $F(x)$.

[21] These statements are not absolutes. It may well be so that an orthodox society will make everyone concede more than a liberal society will do. So the statements

7 Conclusion

This paper has presented a simple theory on 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' disutility when deviating from their bliss points, is more important than the general weight of punishment. In that sense this paper is closely related to the works of Eguia (2011) and Clark & Oswald (1998), who, albeit analyzing different issues than us, do concentrate on how curvature affects individual behavior. To connect the model results to outcomes across societies, and based on empirical and casual observations of punishments in groups and societies, we applied labels to the curvature of social pressure: Orthodox societies are those true to the book and hence utilize concave social pressure; and Liberal societies are those allowing freedom of expression as long as it is not too extreme and hence utilize convex social pressure. These labels are not necessary for the formal analysis, but in our view they are helpful in matching societal traits to actual societies and in producing insights and predictions for these societies.

Plainly, liberal societies facilitate a mentality of compromise, where most individuals are compelled to adjust at least a little bit to the norm. A more intricate result is that the degree of liberalism, i.e., the degree of convexity, plays an important role. Very liberal societies will tend to mainly make those who privately dislike the norm adjust to it fairly much. This will create a society which looks polarized. Less liberal societies will mainly induce those who nearly agree with the norm to make large concessions to it, leading to a unimodal distribution of public stances.[22]

Orthodoxy, on the other hand, facilitates an either-or mentality, since (almost) only full conformity counts. Indeed, this will sometimes lead to full conformity, but may backfire so that some do not concede at all. Moreover, the degree of orthodoxy, i.e. the degree of concavity, is important in predicting who will follow the norm. In very orthodox societies, the full conformers will be those who nearly agree with the norm anyway, while those who strongly reject the norm privately will be alienated also in open. As opposed to that, in less orthodox societies, those who dislike the norm the most will be the only ones upholding it, while those who nearly agree with it in private will pose mild critique of it in public. This creates a surprising result of inversion of opinions.

on who concedes the most really relate to whether the concessions of the extremists are relatively larger or smaller than the concessions of moderates.

[22] The statements regarding distributions of stances hold under sufficiently uniform distribution of types.

In practice it seems plausible that individuals will differ with respect to their degree of perfectionism (or non-perfectionism). Our results should therefore be seen as qualitative statements contingent on a certain relationship between the degree of orthodoxy or liberalism and the degree of perfectionism. For instance, suppose that in a certain society each bliss point is represented by individuals with different degrees of perfectionism. As an orthodox society becomes more orthodox, it will tend to further alienate objectors, while attracting mild supporters to the norm. This is since the share of individuals who are "more perfectionist than society is orthodox" falls. Likewise, as a liberal society becomes more liberal, it will become better at attracting objectors towards the norm, while making it less necessary for mild supporters to conform. This description is of course incomplete, but it shows that nothing in principal prevents the model from being extended or interpreted along this dimension.

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 – skewed norms cannot be sustained in equilibrium. This may be linked to the loose observation that a liberal atmosphere is often correlated with democracy. At the same time, orthodox societies *can* sustain skewed social norms and have rules that do not represent the people's interest, even on average. It can be shown that these results hold and are emphasized if the norm is determined in a median voter framework.[23] In real life, where dynamics play a role, the skewness of the norm in orthodox societies may be materialized as history dependence. That is, the initial set of common rules also determines the long-run equilibrium outcome, even if opinions in society have changed so that the norm is no longer representative. The model further predicts an association between harsh punishments, orthodox societies and extremist norms – only in orthodox societies is it possible to sustain a skewed norm, and the more skewed it is, the harsher is the needed punishment.

References

[1] Asch, S. E., (1955), "Opinions and Social Pressure," *Scientific American*, Vol. 193, No. 5, pp. 31-35.

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

[23] Another alternative would be to relax the basic assumption that a norm exists, and instead let the social pressure arise from the whole distribution of stances. This is substantially more complex and is analyzed in a different paper (Michaeli & Spiro, 2013a,b). There too, we find that skewed norms can arise under orthodoxy.

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

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

[5] Eguia, J.X. (2011). "On the Spatial Representation of Decision Profiles." *Economic Theory*, 2011.

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

[7] Gelfand, M.J., et al. (2011). "Differences between Tight and Loose Cultures: A 33-Nation Study." *Science*, 332, 1100–1104.

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

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

[10] Goffman, E., (1959), Presentation of Self in Everyday Life. New York: The Overlook Press.

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

[12] Hamlin, A., Jennings, C., (2011), "Expressive Political Behaviour, Foundations, Scope and Implications," *British Journal of Political Science*, Vol. 41, No. 3, pp. 645 - 670. DOI: http://dx.doi.org/10.1017/S0007123411000020.

[13] Herrmann, B., Thöni, C., & Gächter, S. (2008). "Antisocial Punishment across Societies." *Science*, 319, 1362–1367.

[14] Holbrook, A. L., Green, M. C. and Krosnick, J. A., (2003) "Telephone Versus Face-to-face Interviewing of National Probability Samples with Long Questionnaires," *Public Opinion Quarterly*, Vol. 67, pp. 79–125

[15] Jones, S. R. G., (1984), The Economics of Conformism. Oxford: Basil Blackwell.

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

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

[18] 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*, 11(3), 495-524.

[19] Kuran, T., (1995a), "The Inevitability of Future Revolutionary Surprises," *The American Journal of Sociology*, Vol. 100, No. 6, pp. 1528-1551.
[20] Kuran, T., & Sandholm, W. H. (2008). "Cultural integration and its discontents". *The Review of Economic Studies*, 75(1), 201-228.
[21] 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.
[22] 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.
[23] Michaeli, M. & Spiro, D., (2013a), "Peer Pressure and Skewed Norms", mimeo, University of Oslo.
[24] Michaeli, M. & Spiro, D., (2013b), "Losing your Religion: The Inversion of revealed preferences under social pressure", mimeo, University of Oslo.
[25] Schelling, T., (1971), "Dynamic Models of Segregation", *Journal of Mathematical Sociology*, Vol. 1, Iss. 2, pp.143–186.

8 Appendix – Proofs and derivations

8.1 Some useful results

8.1.1 Conformity and relative concession

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

$$\frac{ds^*}{dt} = \frac{D''\left(t-s^*\right)}{P''\left(s^*\right)+D''\left(t-s^*\right)} \tag{4}$$

$$\frac{d^2s^*}{dt^2} = \frac{\left[D'''\left(t-s^*\right)\left(P''\left(s^*\right)\right)^2 - P'''\left(s^*\right)\left(D''\left(t-s^*\right)\right)^2\right]}{\left(P''\left(s^*\right)+D''\left(t-s^*\right)\right)^3} \tag{5}$$

Lemma 6 *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 $s^*(t) \in \{\bar{s}, t\}$ 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 4 and 5, 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.* ■

8.1.2 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.
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.

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

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

$$M_i(\tilde{s}^*) = \begin{cases} \frac{t(\tilde{s}^*) - t_i}{t_h - t_l} \text{ if } s^*(t) \text{ is increasing in the subrange } T_i \\ \frac{t_{i+1} - t(\tilde{s}^*)}{t_h - t_l} \text{ if } s^*(t) \text{ is decreasing in the subrange } T_i \end{cases} \quad (6)$$

$$pPDF_{T_i}(\tilde{s}^*) = \frac{dM_i(\tilde{s}^*)}{d\tilde{s}^*} = \frac{1}{t_h - t_l} \left| \frac{dt}{ds^*}|_{\tilde{s}^*} \right| \quad (7)$$

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:

$$\frac{d(pPDF(\tilde{s}^*))}{ds^*} = \begin{cases} \frac{1}{t_h - t_l} \frac{d^2t}{ds^{*2}}|_{\tilde{s}^*} \text{ if } \frac{dt}{ds^*}|_{\tilde{s}^*} > 0 \\ -\frac{1}{t_h - t_l} \frac{d^2t}{ds^{*2}}|_{\tilde{s}^*} \text{if } \frac{dt}{ds^*}|_{\tilde{s}^*} < 0 \end{cases}. \quad (8)$$

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

Lemma 7 *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 8, 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.

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

[25] Note that the previous expressions catch 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 $pPDF's$.

8.2 Proof of Proposition 1

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:

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

$\alpha, \beta > 1$: That every t has a unique inner solution can be easily verified using equations (2) and (3). The statements that $|s^*(t) - \bar{s}|$ is increasing either convexly or concavely follow from applying the implicit function theorem twice to equation (2) 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 6 explicitly for power functions, and substituting the FOC into it. Plugging the expressions for the derivatives of P and D into equation (5), 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 7 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 $\bar{s}$, 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 6 that conformity is decreasing $\forall t \geq \bar{s}$.

$1 = \alpha < \beta$: It is easy to verify that types sufficiently close to the norm declare their type and all types sufficiently far from the norm declare the same stance. For the subrange where all declare their type $ds^*/dt = 1$ and hence $d^2s^*/dt^2 = 0$. For the subrange where all declare the same inner stance $ds^*/dt = 0$ and hence $d^2s^*/dt^2 = 0$. Applying these results to Lemma 6 yields the first three statements of part 2. For a sufficiently large K some types have the same inner solution implying bimodality when the norm is central enough (if it is not central we get a peak only on one side of the norm). For a sufficiently small K all types choose $s^*(t) = t$ implying a uniform distribution.

$1 = \beta < \alpha$: 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 (4) implies $ds^*/dt = 1$ and hence $d^2s^*/dt^2 = 0$. Applying these results to Lemma 6 yields the first three statements of part 1. Since for any $K > 0$ there exists 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:

We will show that if the range of types is broad enough, 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 bi-modal.

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 $L(t, s)$ near the corner solutions.

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

Hence $L'(t, \bar{s}) < 0$ 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 $[\bar{s}, t]$. We will perform the proof only for the case of $\beta > 1$, as when $\beta = 1$ there are only corner solutions, and this case is covered by the proof of parts 1 and 2 of proposition 3.[26] In inner solutions $L'(t, s) = 0$ and hence we get

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

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

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

Since $(t-s)^{\alpha-2} (s-\bar{s})^{-\beta}$ is strictly positive in $]\bar{s}, t[$, and $[(1-\beta)(t-s) - (\alpha-1)(s-\bar{s})]$ is linear in s, negative at $s = \bar{s}$ 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 $]\bar{s}, t[$).

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 $L(t, s)$ has exactly two extrema in the range $]\bar{s}, t[$. From the negative values of $L'(t, s)$ at the edges

[26] That is, the properties of the case $\beta = 1$ and $\alpha < 1$ covered in Proposition 3 are the same as the properties of the limit case when β approaches 1 from above.

of this range we finally conclude that the first extremum (where $f(s)$ is falling) is a minimum point of $L(t,s)$, and the second extremum (where $f(s)$ is rising) is a maximum point of $L(t,s)$. The global minimum of $L(t,s)$ 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 $L(t,s)$ in the range $]\overline{s},t[$, and therefore $s=t$ is the solution to the minimization problem.

Next 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 $L(t_0,s)$ for some t_0, then there exists a local minimum of $L(t,s)$ 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 $\Delta L \equiv L(t,t) - L(t,\tilde{s})$, where $\tilde{s}$ is the stance at which $L(t,s)$ gets the local minimum. Type t prefers the inner solution to the corner solution if and only if ΔL is positive. Thus we need to show that ΔL is negative for small enough $|t-\overline{s}|$ but is increasing in t (and so if ΔL is positive for t_0 it is positive for $t>t_0$ too).

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

and since $\alpha<1\leq\beta$, for small enough $|t-\overline{s}|$ the dominant element is $(t-\tilde{s})^{\alpha}$ and so ΔL is negative (i.e., types close to the norm choose the corner solution of $s^*=t$). Differentiating ΔL 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 4 we have that $\frac{d\tilde{s}}{dt} < 0$ in an inner solution when D is concave, and so $\Delta L_t'' > 0$. Hence ΔL is strictly increasing and strictly convex, implying that for a broad enough range of types, types sufficiently far from 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 $s^* = \bar{t}$, and then it rises at the cutoff where $\Delta L = 0$, and keeps rising as t increases (follows from restating the inequality in part 2 of Lemma 6 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 $\bar{s}$ (if $\bar{s}$ is not sufficiently centered there is only one). 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:

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

Substituting $t - s^* = \left(\frac{\beta K}{\alpha}\right)^{\frac{1}{\alpha-1}} (s^*-\bar{s})^{\frac{\beta-1}{\alpha-1}}$ 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 7 the $pPDF$ for the inner solutions is decreasing towards the edges.■

8.3 Proof of Proposition 2

Lemma 8 *Suppose $\bar{s}$ is the average stance in society and $t \sim U(t_l, t_h)$. Then, for any positive α and β there is an equilibrium where $\bar{s} = \frac{t_h+t_l}{2}$.*

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* $\bar{s} = \frac{t_h + t_l}{2}$, *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* $\bar{s} = \frac{t_h + t_l}{2}$ *can be sustained as a social norm in equilibrium for any values positive of* α *and* β. ■

By Lemma 8, $\bar{s} = \frac{t_h + t_l}{2}$ can be sustained in equilibrium. Hence, we only need to show that $\bar{s} \neq \frac{t_h + t_l}{2}$ cannot be an equilibrium. If $\bar{s} \neq \frac{t_h + t_l}{2}$, then there are types that reside outside the range $[\bar{s} - d, \bar{s} + d]$, all of whom either to the left of $\bar{s}$, such that for each of them $s^*(t) \leq \bar{s}$, or to the right of it (such that for each of them $s^*(t) \geq \bar{s}$). Hence, for $\bar{s}$ to be the average of *all* stances, we must have $s^*(t) = \bar{s}$ for all those types with $|t - \bar{s}| > d$. When $\beta > 1$ all types either have an inner solution or declare their type. Thus follows that there is no equilibrium with $\bar{s} \neq \frac{t_h + t_l}{2}$.

8.4 Proof of Proposition 3

Parts 1 and 2

The second-order condition (equation 3) is positive when $\alpha, \beta < 1$, 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 $|t - \bar{s}| < K^{\frac{1}{\alpha - \beta}}$, and so $s^*(t) = t$ iff $|t - \bar{s}| \geq K^{\frac{1}{\alpha - \beta}}$, and $s^*(t) = \bar{s}$ iff $|t - \bar{s}| < K^{\frac{1}{\alpha - \beta}}$. When $\alpha < \beta$ the converse holds, with $s^*(t) = t$ iff $|t - \bar{s}| < K^{\frac{1}{\alpha - \beta}}$, and $s^*(t) = \bar{s}$ iff $|t - \bar{s}| \geq K^{\frac{1}{\alpha - \beta}}$, which means that conformity is initially decreasing in t but then sharply increases to full conformity at $|t - \bar{s}| = K^{\frac{1}{\alpha - \beta}}$ (where it also stays). The distribution of s^* then follows from this, where the sufficient condition for uniform tails at *both* edges of the distribution in the case of $\beta < \alpha$ is to have types with $|t - \bar{s}| > K^{\frac{1}{\alpha - \beta}}$ at both sides of $\bar{s}$, whereas the sufficient condition for having a peak at $\bar{s}$ in the case of $\alpha < \beta$ is to have types with $|t - \bar{s}| > K^{\frac{1}{\alpha - \beta}}$ at one side of $\bar{s}$. 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 $\alpha < \beta$.

Part 3

We perform the proof for $t \geq \bar{s}$. The opposite case is similar. We also prove only for the case of $\beta < 1$, as when $\beta = 1$ there are only inner

solutions, and this is covered by the proof of parts 1 and 2 of proposition 1.[27] We will prove that if the range of types is broad enough, then: a) types close enough to the norm fully conform, while for types far from the norm $|s^*(t) - \bar{s}|$ is increasing; b) conformity is weakly decreasing in $|t - \bar{s}|$; c) 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 group on each side (provided that $\bar{s}$ is sufficiently centered), peaking at the edge of the range. On 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 $s^* = \bar{s}$, 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 $L(t, s)$ in the edges of this range.

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

Hence $L'(t, \overline{s}) = \infty$ 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 $L'(t, s) = 0$ 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 $f(s) = 0$ in 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:

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

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, 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 $L(t, s)$

[27] That is, the properties of the case $\beta = 1$ and $\alpha > 1$ covered in Proposition 1 are the same as the properties of the limit case when β approaches 1 from below.

has exactly two extrema in the range $]\overline{s}, t[$. From the positive values of $L'(t,s)$ at the edges of this range we finally conclude that the first extremum (where $f(s)$ is rising) is a maximum point of $L(t,s)$, and the second extremum (where $f(s)$ is falling) is a minimum point of $L(t,s)$. The global minimum of $L(t,s)$ 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 $L(t,s)$ in the range $]\overline{s}, t[$, and therefore $s = \overline{s}$ is the solution to the minimization problem.

Next 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 $|s^*(t) - \bar{s}|$ is increasing in t. First notice that $f(s)$ is increasing in t, so if there exists a local minimum of $L(t_0, s)$ for some t_0, then there exists a local minimum of $L(t,s)$ 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 $\Delta L \equiv L(t,\bar{s}) - L(t,\tilde{s})$, where $\tilde{s}$ is the stance at which $L(t,s)$ gets the local minimum. Type t prefers the inner solution to the corner solution if and only if ΔL is positive. Thus we need to show that ΔL is negative for small enough $|t - \bar{s}|$ but is increasing in t (and so if ΔL is positive for t_0 it is positive for t_1 too).

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

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

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

Using the first order condition

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

Differentiating once more

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

By equation 4 we have that $\frac{d\tilde{s}}{dt} > 1$ in an inner solution when P is concave, and so $\Delta L''_t > 0$. Hence ΔL is strictly increasing and strictly convex, implying that for a broad enough range of types, 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.

b) The statement on conformity follows directly from this.

c) By the definition of relative concession it equals 1 at the range of types choosing $s^* = \bar{s}$, and then it falls at the cutoff where $\Delta L = 0$, and keeps falling as t increases (follows from restating the inequality in part 2 of Lemma 6 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 $\bar{s}$. Otherwise, if the range of types is broad enough (so that some have an inner solution), the resultant distribution of stances has a peak at $s = \bar{s}$ with a tail at each side of it (if $\bar{s}$ is not sufficiently centered there is only one tail). The tails are detached since for the type t who is indifferent between the corner and inner solution the inner solution s^* is necessarily strictly greater than $\bar{s}$ (because $L'(t, \bar{s}) = \infty$ while $L'(t, s^*) = 0$ 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 1 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 7 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.■

8.5 Proof of Proposition 4

Lemma 9 *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 $\bar{s} \neq \frac{t_h+t_l}{2}$ 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 $\bar{s} > \frac{t_h+t_l}{2}$, then $[\bar{s} - d, \bar{s} + d] \subset [t_l, t_h]$ and $\forall t \in [t_l, t_h] \setminus [\bar{s} - d, \bar{s} + d]$ we have $s^*(t) \leq \bar{s}$. For $\bar{s}$ to be the average of all stances, it is then necessary that $s^*(t) = \bar{s}$ $\forall t < \bar{s} - d$.* ■

Lemma 8 ensures that $\bar{s} = \frac{t_h+t_l}{2}$ is a possible equilibrium. Furthermore:

1. If $s^*(t) = \bar{s}\ \forall t$, then $\bar{s}$ is the average of all stances, regardless of its value. This concludes sufficiency. Suppose that not all types state the norm and that $\bar{s} > \frac{t_h+t_l}{2}$. Then, from Proposition 3 part (3), we know that types who do not fully conform will have an inner solution. Then, when $\bar{s} > \frac{t_h+t_l}{2}$, $s^*(t_l) \neq \bar{s}$, which by Lemma 9 implies that $\bar{s}$ cannot be sustained in equilibrium. A corresponding argument applies to $\bar{s} < \frac{t_h+t_l}{2}$. This concludes necessity.

2. From Proposition 3 part (1) we know that types with $|t - \bar{s}| < \Delta$ choose $\bar{s}$ as their stance. If $t_h - t_l < 2\Delta$, then the condition $\bar{s} \in [t_h - \Delta, t_l + \Delta]$ implies that $\bar{s} - t_l < \Delta$ and $t_h - \bar{s} < \Delta$. This ensures that every type in $[t_l, t_h]$ chooses $\bar{s}$ as a stance, and so the distribution of stance is degenerate at $\bar{s}$, which implies that $\bar{s}$ is the average stance. This concludes sufficiency. Otherwise, suppose $t_h - t_l > 2\Delta$. Then $[t_h - \Delta, t_l + \Delta]$ is an empty set. Alternatively, suppose $\bar{s} \notin [t_h - \Delta, t_l + \Delta]$. Then either $s^*(t_l) \neq \bar{s}$ or $s^*(t_h) \neq \bar{s}$, which by Lemma 9 implies that $\bar{s}$ cannot be sustained in equilibrium. This concludes necessity.

3. 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 $t_h - t_l > 2\Delta$, then the condition $\bar{s} \in [t_l + \Delta, t_h - \Delta]$ implies that $\bar{s} - t_l > \Delta$ and $t_h - \bar{s} > \Delta$, and so ensures that all the 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. Alternatively, 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 9 implies that $\bar{s}$ cannot be sustained in equilibrium. A corresponding argument applies to $\bar{s} < t_l + \Delta$. This concludes necessity.■