Opinion Expressions under Social Sanctions

Mehmet Bac*

December 18, 2013

Abstract

I study a social debate where individuals are subject to informal sanctions if their expressions or silence signal the opinions of a minority group. Individual preferences are peaked at the expression of true opinions and there is a loss of utility from keeping silent. The model generates predictions about how equilibrium expressions change as a function of model primitives such as sanction intensity, disutility of silence and size of the minority group. A dynamic extension sheds light on the limit distribution of opinions if unvoiced opinions gradually disappear while publicly expressed opinions gain new adherents over time.

JEL Classification Numbers: D78, D72, Z13.

Key Words: Opinion expression, social sanctions, norms, Bayesian equilibrium.

*Sabanci University, Faculty of Arts and Social Sciences, Tuzla, Istanbul 34956, Turkey. E-mail: bac@sabanciuniv.edu

1 Introduction

In any society and time certain ideologies in politics, beliefs in religion, styles in arts, clothing and family life are considered unacceptable, or simply out of fashion. These norms deter deviants by legitimizing informal social sanctions that range from withdrawing sympathy and support to outright violence. As they vary in form and intensity, social sanctions can have important consequences. The fear of evoking scrutiny and criticism can shout out opinion expressions, leave unchecked the extreme variants of the dominant majority and, potentially, homogenize expressions. While diversity of expressions is generally praised for conducing productive social debates and better choices, circumstances exist in which censorship of certain types or forms of expressions might be beneficial−for example, silencing individuals who praise vandalism or terrorism. It is therefore important to identify the characteristics of media that favor anti-speech norms and strengthen social censorship, to improve our understanding as to why and whose expressions are likely to be distorted in a given social debate.

The static and dynamic effects of social sanctions on expressions is a live research area in political science, sociology and allied disciplines.[1] The related literature can be classified broadly in two categories. The first line of research begins with the works of Schelling (1978) and Granovetter (1978) and applies critical mass models to study the social consequences of individual choices in topics such as collective action problems, voting, bank runs, and even revolutions.[2] The second line of research is a vast and growing public opinion and communication literature, based on Noelle-Neumann's (1974) *spiral of silence* theory of public opinion formation.[3]

[1]The social sanctions targeting specific opinion groups can be complex and rooted in history; in some cases they are strategically nourished by political speech, upon citizens' demand. Glaeser (2005) provides an interesting account and analysis of hatred, strategically supplied and demanded at the group level. The study of the mechanisms by which social sanctions are applied is beyond the scope of this paper.

[2]These are discrete choice models with heterogeneous agents whose individual payoffs increase when others behave similarly. Over the past few decades they have been extended in several directions to study conformism, path-dependence of collective choices and related phenomena; examples include Akerlof (1980), Jones (1984), Kuran (1987), Olivier et al (1985), Bernheim (1994) and Chwe (1999). Brock and Durlauf (2001) develop a generalized critical mass model with microfoundations. More recently, Benabou and Tirole (2006) offer an analysis of optimal incentive provision under pressures to conform in a continuum-agent model.

[3]According to this view, the power of the majority to threaten minority expressions serves to achieve and institutionalize consensus. Noelle-Neumann (1974, 1993) posits that individuals keep

Interdisciplinary and rich in ideas, this literature develops and tests hypotheses about determinants of public expression outcomes. It lacks, however, formal models based on explicit individual motives and choice, capable of generating a rich set of expression phenomena as equilibrium outcomes.

This paper develops a linear model of opinion expressions, similar in spirit to those of social conformity in the tradition of Bernheim (1994). Individuals can express any opinion of their choice or they can remain silent. The motivation to express an opinion is associated with an expressive utility, peaked at expression of own opinion, whereas silence produces a psychic cost or a loss of integrity relative to freely expressing one's own opinion. Given a profile of expressions, individuals commonly perceived to hold the minority opinions which the orthodox majority considers intolerable can be subject to informal sanctions.[4] The sanction per victim is assumed to be a decreasing function of the minority population. With these ingredients, the model delivers predictions about expression strategies and inferred opinions of the individuals. Who expresses what, who the silent, who the vocal and who the sanctioned are depend on the sanction intensity, the cost of silence, individual preferences over expressions and the relative size of the minority. I reformulate the results by linking the model's parameters to observable characteristics of debates and expression media. Finally, I comment on the model's implications regarding the evolution of the true opinion distribution under a reasonable assumption about the influence of public expressions on true opinions.

The model borrows elements from the continuum-agent models in Kuran (1987) and Dharmapala and McAdams (2005). In Kuran, agents motivated by reputational utility express one of the two extreme positions and determine the public opinion. Kuran is interested in conditions of policy continuation and sudden drastic shifts

silent or conform when they perceive a climate of opinion that is hostile to their own viewpoint, lest they experience the negative consequences of supporting unpopular opinions. Experimental studies confirm the fear of isolation and sanctions in social settings. In Hayes et al. (2000), for example, when asked to select from a list of topics for discussion, subjects displayed great preference for a particular topic when their own opinion was more consistent with the popular opinion. See Scheufele and Moy (2000) for a critical evaluation of the extensive empirical literature.

[4]Those who express specific opinions are punished for what they think or believe, not for what they express. One justification for this approach is that preferences or types, not present acts, determine future actions. For example, a speaker who reveals an extreme racist position may be subject to social sanctions because his type is taken as an indicator of his future behavior; the arguments in the speech are relevant to the extent that they correctly signal the type of the speaker.

in public support that follow minor shocks, whereas Dharmapala and McAdams focus on the impact of formal and informal sanctions on crimes induced by hate speech. With respect to these works, the emphasis here is on the magnitude of distortions in expression strategies, explaining which opinions are absent, who speaks up and who is silenced on the opinion spectrum when social sanctions depend on the relative size of the target minority. Recent studies have demonstrated the relevance of silence when individuals experience a fear of isolation in expression media, including survey interviews where social pressures are considered minimal. As I show, the dynamics of equilibria with silence differ from those without silence because distorted conforming voices more than silence can contribute to the growth of conformism.

To highlight some of the model's predictions, in equilibrium sanctions can be ineffective on the minority if individuals perceive a large cost from remaining silent. The first whose strategies are to be affected by social sanctions and the first to disappear from public expressions are the majority neighbors of the minority, not the target minority group itself. Generally, the set of sanctioned expressions is never confined to the minority range−majority opinions that come sufficiently close to the minority are also sanctioned. A small dose of informal sanction generates silenced opinions by inducing the majority neighbors to distance themselves from the minority, thus building a gap between the expressions of the two camps. Public opinion scholars associate such small sanctions with "descriptive" or informational social norms. An example would be the mild social disapproval for expressions against recycling policies.[5] On the other hand, in an environment in which the social sanction and the cost of silence are both large, social pressures to conform are powerful, so, opinion misrepresentation is common and many opinions are absent. If the social sanction is large yet remains smaller than the cost of silence, I show that equilibrium expressions may even display a greater variety of minority opinions than majority opinions.

As Harrison (1940) argued long ago and these equilibria confirm, expressions are not exactly what people think, but what people are willing to publicly acknowledge they think. In this model the distribution of true opinions and the distribution of expressed opinions never coincide under positive social sanctions. Factors that lead to increases in the social sanction widen the expression gap between the two

[5]See Lapinski and Rimal (2005) for a discussion and typology of social norms and informal sanctions.

groups because majority members increasingly misrepresent their opinions to distance themselves from the minority. The model predicts that combinations of small minority, large sanction intensity and large cost of silence lead to full conformity of expressions with the orthodox majority views located at the opposite extreme of the minority. This could be the case in morally loaded debates like those involving racial politics where individuals feel strong pressures to moderate their racially conservative views.

The question as to when silence is preferred to some form of expression as an equilibrium individual strategy is interesting. The answer of course depends on social beliefs. For instance, if silent individuals are always inferred as minority members and subject to social sanctions, no individual would remain silent because truthful expression of own opinions would dominate silence, as remaining silent entails a psychic cost in addition. To discard equilibria in which individuals are artificially forced to express an opinion by fear of sanctions on off-the-equilibrium silence, I impose a "right to silence" condition on beliefs about the types of silent individuals. This condition allows majority members to become silent if they wish so, without fear of social sanctions. I show that under this condition in equilibrium a silent group always consists of the entire minority plus a range of majority neighbors. In other words, silence, though individually costly, becomes a sanction-free pooling outcome. Besides the belief system, the other key determinant of the identities of silent and vocal individuals is the magnitude of the social sanction relative to the cost of silence. Equilibria with silence emerge in debates involving a small minority and/or large sanction intensity, provided the individual cost of silence is not too large. An example to this kind of environment is opinion polls on socially difficult issues that touch upon punitive norms, in a medium of expression where "involvement obligation" is small, hence the cost of silence is small. If the silent respondents are wrongfully interpreted as indifferent or lacking an opinion while they overwhelmingly hold similar opinions which they prefer to hide, the resulting measure of public sentiment may miss a significant base and lead to an overstatement of support for a specific public action.

Finally, in a dynamic extension of the model I investigate the evolution of true opinions under the assumption that silenced opinions lose support, i.e., density, to voiced opinions. Assuming such a process at work, the model produces a rich set of possible evolutions of the true opinion distribution. In one of these, the minority group grows to the detriment of the majority and the sanction per victim

diminishes over time. I argue that this is plausible in a tolerant society debating a morally loaded issue, or, expressed in terms of the model's parameters, under a large cost of silence relative to the social sanction. On the other hand, there are many circumstances in which social sanctions eventually lead some or all minority members to switch to the majority side. Minorities are likely to keep loosing their adherents if they are silent in the initial equilibrium. I illustrate some of these dynamics with empirical findings from the literature on the evolution of public opinion on same-sex marriage, abortion and school integration issues.

2 The Model

A society consists of a continuum of individuals, distributed on the unit interval according to their opinions on a given issue. Individual $s \in [0,1]$ is of the opinion, or type, s. The unit interval can be interpreted as the range of positions on social issues such as human rights, race, terrorism, immigration, admission of religious symbols in the education system or conformity with a dressing code in public, etc. Let $\Gamma(.)$ denote the cumulative distribution function of individual opinions, $g(.)$ the corresponding density function and $\hat{s}$, the median opinion. Each individual's opinion is private, but the distribution of opinions is publicly known. A borderline opinion $\gamma \in (\hat{s}, 1)$ separates the society in two camps, such that opinions in the range $[\gamma, 1]$ are in minority. The analysis admits all $\gamma > \hat{s}$ and delivers a predicted outcome of expressions for any range of minority opinions.[6]

The individual *expression strategy*, $v : [0,1] \to [0,1] \times \emptyset$, assigns an opinion from $[0,1]$ or silence, $\{\emptyset\}$. Then an *expression outcome* is a profile of expression strategies $\{v_s\}_{s\in[0,1]}$, i.e., a collection of expressed opinions plus a group of silent individuals. Expression outcomes are publicly observable.

Individuals' utility functions are made up of two components. The *expressive utility* component represents the satisfaction associated purely with expression of an opinion, or, if no opinion is expressed, the sacrifice of integrity from self-censoring. The second component is the *sanction disutility*, experienced only if a social sanction

[6]The establishment of new expression groups which could isolate themselves from the rest of the population, or secession of the minority to form a relatively homogenized population on its own, is ruled out. Although this simplification suppresses the possibility that individuals otherwise subject to social sanctions in the broader media can derive some expressive utility by freely voicing their opinions within their groups, typically many issues inevitably require some engagement with the rest of the population.

is imposed on the individual. While expressive utility promotes independence and truthful expressions, sanction disutility generates a pressure to conform or keep silent. The utility of individual s is

$$U_s = U_s^E - \iota(v)f,$$

where f denotes the social sanction and $\iota(v)$ is an indicator function such that $\iota(v) = 1$ if and only if the expression v triggers the sanction. I adopt a simple form for expressive utility:

$$U_s^E = \begin{cases} -|v - s| \text{ if } v \in [0,1]; \\ -\alpha \text{ if } v = \emptyset. \end{cases}$$

The expressive utility of individual s is single-peaked at s and silence generates the disutility $\alpha > 0$.[7]

A social sanction is imposed by individuals from the majority group on individuals who are commonly perceived to hold a minority opinion. Thus, the sanction is triggered by common perceptions and targets individuals for their true opinions or internally held preferences and attitudes. Observed expressions serve as signals to form these perceptions. As such the social sanction contrasts with the formal sanctions confined exclusively on expressions of specific opinions. Perception-based social sanctions also produce richer sets of outcomes than pure formal sanctions. In all equilibria of this model minority expressions are sanctioned, but the range of sanctioned expressions is larger than the range of minority opinions.

Perceptions about individuals are represented by a common belief system. Given a profile of expressions $\{v_s\}_{s\in[0,1]}$, a common belief system assigns to each expression v a probability $\mu(v|\{v\})$ that the individual s who expresses v holds a minority opinion; that is, $\mu(v|\{v\}) = prob(s \geq \gamma|\{v\})$.[8] I allow an individual to avoid the sanction if he is perceived to be a minority with probability less than one, that is,

[7] $|x|$ denotes the absolute value of x. Results go through under more general symmetric and single-peaked expressive utility functions−an example is the quadratic form, $-(v-s)^2$. See Kuran (1995) pp. 30-35 for a formal discussion of expressive utility. The assumptions of costless sanction enforcement and common cost of silence are both motivated by simplicity of exposition. I comment on the impact of heterogeneous costs of silence in the concluding section.

[8] Because all individuals share common prior beliefs and observe the entire range of expressions, in equilibrium they will hold common posterior beliefs on the types of individuals who actually express an opinion.

$\iota(v) = 1$ only if $\mu(v|\{v\}) = 1$.[9]

The last component of the model is determination of the sanction size. The model does not explicitly incorporate the micro-level processes through which the sanction is imposed; it does not aim at explaining who among the majority impose the sanction and contribute to its intensity. The sanction building process would inhibit several complexities: Participation to punishment and norm enforcement would vary according to personal traits, occupation and social position of each individual (which requires introduction of a second type dimension for the personal benefit from punishing minorities) and each majority member would have an incentive to free ride on others' punishment efforts (suggesting a public good game structure for the supply of sanctions.) Instead of going into this terse modeling exercise, I assume that the per-victim size of the majority-imposed sanction increases in the relative size of the majority. The specification below distinguishes between two cases: the case where the sanction is actually imposed on commonly identified minority members and the case where it is not.

$$f = \begin{cases} \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1), \text{ if a sanction is imposed in equilibrium,} \\ F > 0, \text{ otherwise.} \end{cases} \tag{1}$$

When minorities are correctly identified and thus sanctioned by the majority, relative group size affects the per-victim sanction as stated in the first line of (1). In this case the sanction per victim declines in the relative size of the minority, approaching zero as the group sizes converge to each other.[10] The intensity parameter $\kappa > 0$ represents determinants of the sanction other than the relative group size. The sanction form $f = \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1)$ is chosen here for its convenience; alternative specifications are admissible. The second line of (1) applies when no individual is

[9]The equilibria presented in Section 3 are not affected if social sanctions are imposed for beliefs $\mu \in (\overline{\mu}, 1]$ where $0 < \overline{\mu} < 1$. A smaller $\overline{\mu}$ points to a less tolerant majority and as such it only restricts the possibility that a group of minority and majority members pool at the same expression. See the discussion following introduction of the equilibrium concept and Appendix A.

[10]The relative group size effect can be found in the writings of David Hume, John Locke and Jean-Jacques Rousseau−some of which are quoted in Noelle-Neumann (1979). James Madison (1961), for instance, writes: the "practical influence [of each individual's opinion] on his conduct depends much on the number which he supposes to have entertained the same opinion." This influence on expressions operates through potential inter-group pressures and sanctions. The functional form in (1) is in the same spirit as Kuran's (1987) assumption that an individual's benefit from complying with an extreme opinion is proportional to the "vote share" of that opinion.

identified as minority, hence, when no individual is actually sanctioned. Now the sanction denoted by F rather serves as a threat on an individual who may deviate to a range of "off-the-equilibrium," unvoiced, opinions (which includes, but may not be confined to, the minority range.) The size of the sanction applied to a single potential deviant is likely to be different from the case where a continuum of minority members are sanctioned by the majority. Lacking a justification for an assumption concerning the size of an absent sanction, F is treated as a parameter; its determination and the expression range on which its threat is pending are part of the equilibrium construction exercise.

The sequence of events is as follows. Individuals simultaneously determine and execute their expression strategies. Based on observed expressions, beliefs about each individual's true opinion are formed. Finally, sanctions (if any) are applied and individual utilities are realized.

An *expressions equilibrium* $(\{v^*\}, \mu, f)$ is a collection of expression strategies, a belief system and a social sanction such that expression strategies are individually optimal given μ and f, while the belief system is consistent with the strategies and f is determined by (1). The belief system satisfies two additional conditions:

B1. Consider an equilibrium in which there exist *vocal* minority members. If an opinion $t < \gamma$ is not expressed and $U^*_\gamma \geq -|\gamma - t|$, then $\mu(t|\{v^*\}) < 1$.

B2. Consider an equilibrium in which silence is sanctioned, i.e., $\mu(\emptyset|\{v^*\}) = 1$. If the equilibrium strategy of a majority member $s < \gamma$ is $v^* \in [0, 1]$, then $|v^* - s| < \alpha$.

I also adopt a tie-breaking convention: If an individual is indifferent between silence and expressing an opinion $s \in [0, 1]$, then he expresses s. Indifference between expressing own sanctioned opinion and a different but sanction-free opinion is broken in favor of expressing own opinion.

Conditions B1 and B2 isolate equilibria that can be supported by somewhat strange beliefs, such as those in which all individuals express exactly the same opinion $0 < t < \gamma$ supported by beliefs that any other expression must come from minority members. B1 is in the spirit of the *Intuitive Criterion* proposed by Cho and Kreps (1987). If a deviant individual expresses an unvoiced majority opinion t which even the borderline minority member γ cannot beneficially imitate, B1 rules out the inference that the deviant is a minority member. The belief system should assign $\mu(t|\{v^*\}) < 1$ and expression of the opinion t should be sanction-free. This condition isolates equilibria in which a majority member is induced to keep silent

or misrepresent his opinion despite the fact that his opinion cannot beneficially be mimicked by a minority member.

Condition B2 essentially restricts beliefs about the types of silent individuals. It would be atypical for a majority to punish its own silent members who prefer silence over their actual expression strategies. B2 rules out this possibility and allows majority members to deviate to silence without the fear of a sanction. While it confers a "right to silence" to the majority, B2 does not provide a safe haven for the minority; silent or vocal, the latter is subject to sanctions whenever correctly identified. However, in any equilibrium in which a positive measure of majority members remain silent, Bayes' rule implies $\mu(\emptyset, \{v^*\}) < 1$ and silence becomes a sanction-free deviation option for all, including, thus, minority members.

The analysis sets aside two types of equilibria. The first type involves equilibria with an all-silent population, which is unnatural in this setting for it means that the majority silences itself by its own threat to sanction any opinion expression.[11] The second type of equilibria owe their existence to the assumption linking perceptions to sanctions, namely, that no sanction is imposed unless one is commonly perceived to be of the minority with probability $\mu = 1$. As a result, one can construct equilibria in which an appropriate mixture of minorities and majorities express the same opinion, generate the belief $\mu < 1$ and avoid the sanction. The structure of these equilibria is shown in Appendix A and shall not be considered in Section 3 for the sake of briefness.[12]

I close this section by relating three parameters of the model, α, κ and γ, to concepts and variables which an interdisciplinary body of research identifies as relevant and operational. Table 1 summarizes the expected qualitative relations. The discussion puts the results in perspective and is useful for future empirical work.

• **The cost of silence, α.**

A measure of the feelings of shame and loss of integrity from renouncing the

[11] Of course, this is not to claim that a silent population cannot ever be a reasonable equilibrium outcome in another setting. For example, such an outcome would arise quite naturally in the presence of a state-imposed large formal sanction on the entire range of expressions. I do not consider state censorship activities in this paper.

[12] In addition, in all these equilibria the size of the interval of unvoiced opinions is identical to the one in Proposition 1, where minorities and majorities separate and the group to which an individual belongs (if not his exact opinion) is fully revealed.

right to expression, α can be linked to at least three factors: issue relevance, issue awareness and the medium of expression.[13]

(i) *Issue relevance.* The cost of silence is large in highly controversial, morally loaded discussions.[14]

(ii) *Issue awareness.* The better the quality and quantity of publicly available information on the issue, the stronger is the basis to form an opinion. Then a general issue awareness feeds confidence in opinions, strengthens the incentives to express opinions and thus raises the cost of remaining silent. Note also that it is natural to expect a correlation between issue awareness and issue relevance.

(iii) *Medium of expression.* The pressure to express an opinion, or individuals' *involvement obligation*, depends on the characteristics of the medium of expression. The cost of silence is large in face-to-face social interactions where individuals are likely to perceive a duty to defend a position and/or fear that their views are isolated. In contrast, involvement obligation is low in mediated internet chatrooms where social presence and contacts are minimal; anonymity of expressions reduces the cost of silence and offers the participants greater latitude to express extreme opinions.[15]

• **The sanction intensity parameter κ.** This parameter represents the majority's degree of disapproval associated with the minority opinions. It thus captures a broad set of factors that may affect the per-victim sanction including the majority's ability and willingness to punish. The willingness to punish is associated with social and economic factors such as permissiveness towards unorthodox expressions, the likelihood of involvement in exchange with other citizens and the potential benefits

[13]Scholars have recently attempted to measure α through survey methods. Hayes et al. (2005) construct a "willingness to self-censor" scale by aggregating respondents' levels of agreement with eight statements including "It is difficult for me to express my opinion if I think others won't agree with what I say," and "There have been many times when I thought others around me were wrong but I didn't let them know." A high score on this scale is considered an indication of strong willingness to self-censor.

[14]Perceptions about the relevance of an issue depend on intrinsic variables that are rooted in individuals' preferences (one may not voice an opinion simply because one does not care) as well as extrinsic variables that are related to the impact expected from opinion expression (the cost of silence falls when people believe expression will have no positive consequence at all.) The number of individuals one can reach by speaking up should thus affect the cost of silence, but this is an attribute of the medium of expression.

[15]See McDevitt et al (2003). Ho and McLeod's (2008) experimental results also indicate a large cost of silence in face-to-face interactions relative to computer-mediated communications.

Table 1: Linking attributes of expression environments to model parameters ("$\sim$" indicates ambiguous sign effect.)

Higher:	α	κ	$[\gamma, 1]$
issue awareness, issue relevance	$+$	$+$	$\sim$
involvement obligation induced by the medium	$++$	$+$	$\sim$
social/economic interdependence between individuals	$\sim$	$-$	$\sim$
tolerance of dissenting opinions	$+$	$--$	$-$

from these interactions. Expect a small κ in open societies. Expect also a small κ in societies that institute strong networks that develop and deepen social and economic interdependence.[16] As for the majority's ability to punish, it depends on the medium in which opinions are expressed: expect a large κ in face-to-face interactions where social presence and involvement obligation are high. Indeed, issues and media that are characterized by a large cost of silence, more likely than not, also involve a relatively large sanction intensity. As α, κ should be positively associated with issue relevance and issue awareness.[17]

• **The range of minority opinions, $[\gamma, 1]$.**

The location of the borderline individual γ, beyond which lies the group of individuals whose opinions the majority does not tolerate, depends on the issue, context and cultural attributes of the society. To illustrate, where opinions about the scope of an ethnic minority's rights vary from the nationalist discourse ($s = 0$) to the extreme secessionist favoring violence ($s = 1$), the actual borderline of tolerable opinions could be $\gamma_1 =$"no more than the right to press and broadcasting in own language" or $\gamma_2 =$"switch to a federalist system with administrative autonomy for the ethnic group." The location of γ may change in time: while the majority could tolerate the expression of γ_2 in peace, in a period of external conflict even γ_1 may not be tolerated. As I show below, the expression equilibria can look quite different

[16]Glaeser's (2005) model links this attribute to the supply of group-level hatred−a fundamental cause for social sanctions. See also Lazear (1999) for a formal approach which identifies the power of economic ties as a basic stimulus for cultural exchange, increasing, therefore, social tolerance of diverse opinions.

[17]It should be noted that these two parameters do not always move together; for example, a strong and widely shared culture that gives a respectful hearing to different opinions has a small κ but a relatively large α.

in two societies that differ substantially in γ values, given identical κ, α and opinion distribution $\Gamma(.)$.

3 Expressions Equilibria

The first part of this section presents some definitions and elementary results. Because their opinions are sanctioned in any equilibrium, minority members consider alternative sanction-free strategies such as expressing a majority opinion or remaining silent. Among the minority the borderline individual at γ has the smallest cost of imitating a given majority opinion, so his behavior plays an important role throughout the analysis. This is reflected in the definitions of two critical opinions, s_c and s_s (see Figure 1.)

Definition 1

$$s_c = \gamma - \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1); \qquad s_s = \gamma - \alpha.$$

[Figure 1]

The opinion s_c is such that, given the social sanction f, the borderline individual can go a maximum distance of $\gamma - s_c$ in imitating the sanction-free majority opinions. He prefers expressing his own sanctioned opinion to the majority opinions in the range $[0, s_c)$. Any factor that increases the per-victim social sanction induces the borderline individual to accept a larger sacrifice of expressive utility and leads to a fall in s_c. On the other hand, if silence is a shelter from the social sanction, no minority member would sacrifice more than α to express a sanction-free majority opinion; the farthest majority opinion that the borderline individual γ would voice is $s_s = \gamma - \alpha$. The following properties are easily verified.

Lemma 1 *s_c is decreasing in κ, increasing in γ if and only if $[1-\Gamma(\gamma)]^2 > \kappa g(\gamma)$ and admits an interior maximum in γ.*

Whereas the impact of κ on s_c should be expected, that of γ may not be so. Given the social sanction, hence, the maximum expressive utility that the borderline individual would sacrifice, an increase in γ will raise s_c by the same amount. But there is a second effect which works in the opposite direction, through the impact of γ on the social sanction. A marginal increase in γ raises f by $\kappa g(\gamma)/[1-\Gamma(\gamma)]^2$,

producing the total effect $1 - \kappa g(\gamma)/[1 - \Gamma(\gamma)]^2$ on s_c. For γ sufficiently large, the second effect dominates the first and the total effect becomes unambiguously negative. So, when the minority is very small and the per-victim sanction is very large to begin with, further reductions in the minority size decrease s_c, which eventually becomes zero when the minority range is reduced to $[\gamma_0, 1]$.

Definition 2 *$\gamma_0(\kappa)$ satisfies $\gamma_0 - \kappa(\frac{\Gamma(\gamma_0)}{1-\Gamma(\gamma_0)} - 1) = 0$, that is, $s_c = 0$.*

In this model the size of the minority group affects its members' incentives to keep silent and misrepresent their opinions. Below I introduce a critical minority size $[\underline{\gamma}, 1]$ such that the borderline individual $\underline{\gamma}$ becomes indifferent between three options: silence, expressing his own sanctioned opinion, and expressing the sanction-free majority opinion s_c (opinion misrepresentation.) If γ is smaller than $\underline{\gamma}$, the social sanction falls below the cost of silence and the borderline individual prefers expressing s_c to silence; the opposite holds for $\gamma > \underline{\gamma}$.

Definition 3 *$\underline{\gamma}(\kappa, \alpha)$ is a critical γ such that $s_c = s_s$, that is, $\alpha = \kappa[\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1]$.*

Note that $\underline{\gamma}(\kappa, \alpha)$ is well-defined.[18] Because characterization of equilibria is carried out mainly in terms of the range of minority opinions, it is useful to identify the condition determining the relative positions of the two critical γ values, $\underline{\gamma}$ and γ_0. Recall that γ_0 is at the same time a critical sanction level such that the borderline individual is wiling to voice the extreme orthodox majority opinion $s = 0$ to avoid the sanction. So, if α is larger than γ_0, which according to Lemma 2 happens if $\underline{\gamma} > \gamma_0$, remaining silent will be a strictly dominated option for the borderline individual.

Lemma 2 *(i) $\underline{\gamma}(\kappa, \alpha) < \gamma_0(\kappa)$ if and only if $\gamma_0(\kappa) > \alpha$.*

(ii) $\underline{\gamma}(\kappa, \alpha)$ is decreasing in κ and increasing in α; $\gamma_0(\kappa)$ is decreasing in κ.

Consider now the case where silence is a strictly dominated option for the borderline individual (the case $\alpha \in (\gamma_0(\kappa), 1]$) despite a large per-victim social sanction (small minority, $\gamma > \gamma_0(\kappa)$.) Such a small minority splits into two groups: those

[18] To see this, let $q(\gamma, \kappa, \alpha) = s_c - s_s$. By definition, $q(\underline{\gamma}(\kappa, \alpha), \kappa, \alpha) = 0$. Note that $q(\gamma, \kappa, \alpha) > 0$ as $\gamma \to \hat{s}$ from above, and $q(\gamma, \kappa, \alpha) < 0$ as $\gamma \to 1$ (see Figure 1.) Since $\frac{\partial q(.)}{\partial \gamma} = -\kappa g(\gamma)/(1 - \Gamma(\gamma))^2 < 0$ and is continuous in γ, $q(.)$ is monotonically decreasing and continuous in γ. It follows that $\underline{\gamma}(\kappa, \alpha) \in (\hat{s}, 1)$ is unique.

holding opinions in the range $[s_\gamma, 1]$ who prefer expressing their own sanctioned opinions to the sanction-free opinion $s = 0$, and those whose opinions are in $[\gamma, s_\gamma)$ and prefer the opposite.

Definition 4 *For $\alpha \in (\gamma_0(\kappa), 1]$ and $\gamma > \gamma_0(\kappa)$, let* $s_\gamma = min\{1, \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1)\}$.

Armed with these definitions, the analysis proceeds with characterization of expressions equilibria, beginning with equilibria in which all individuals speak (type-1), followed by those in which some individuals choose silence (type-2).

3.1 Vocal Equilibria

The expressions game does not admit a fully separating equilibrium in which all individuals express their true opinions unless the social sanction is zero, i.e., unless $\kappa = 0$.[19] All equilibria involve a mixture of separation and pooling in either silence or expression of a specific opinion, moreover the social sanction is never confined to minority expressions.

In *type-1* equilibria, every individual voices an opinion. These equilibria arise when the cost of silence exceeds the social sanction, which corresponds to a small sanction intensity parameter κ and a large minority. Proposition 1 further distinguishes between equilibria in part (i) where all minorities voice their own opinions, are correctly identified and sanctioned, and part (ii) equilibria which involve a positive measure of minority members complying with the expression $v = 0$, along with the entire majority.[20]

Proposition 1 *(i) If $\gamma \in (\hat{s}, min\{\gamma_0(\kappa), \underline{\gamma}(\kappa, \alpha)\}]$, the expressions equilibrium is:*

$$\textit{Strategies:} \quad v_s^* = \begin{cases} s, & \textit{if } s \leq s_c \textit{ or } s \geq \gamma, \\ s_c, & \textit{if } s \in (s_c, \gamma); \end{cases} \tag{2}$$

[19] The intuition is simple: in an equilibrium in which $v_s = s$ for all $s \in [0, 1]$, the belief system satisfies $\mu(s|\{v\}) = 1$ if and only if $s \geq \gamma$, implying the utility $-f$ for individuals $s \geq \gamma$, 0 for individuals $s < \gamma$. Clearly, given any $f > 0$, the borderline individual γ will beneficially deviate from $v_\gamma = \gamma$ to $v = \gamma - \epsilon$ for ϵ arbitrarily small.

[20] Proposition 1 refines the first two propositions in Dharmapala and McAdams (2005). Equilibria are characterized below in terms of the minority size, the social sanction parameter and the cost of silence, as opposed to a fixed social sanction where all individuals are required to express an opinion (silence is not allowed.)

$$\textit{Belief system:} \quad \mu(v_s|\{v^*\}) = \begin{cases} 1, \ \textit{if } v_s > s_c, \\ 0, \ \textit{if } v_s \leq s_c, \\ \in [0,1] \ \textit{if } v_s = \emptyset \end{cases}$$

(ii) If $\gamma_0(\kappa) < \underline{\gamma}(\kappa, \alpha)$ *and* $\gamma \in (\gamma_0(\kappa), min\{\alpha, 1\})$*, the expressions equilibrium is:*

$$\textit{Strategies:} \quad v_s^* = \begin{cases} s, \ \textit{if } s \geq s_\gamma, \\ 0, \ \textit{if } s < s_\gamma; \end{cases} \tag{3}$$

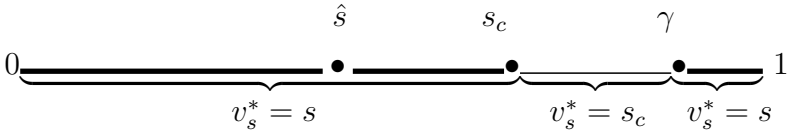
$$\textit{Belief system:} \quad \mu(v_s|\{v^*\}) = \begin{cases} 1, \ \textit{if } v_s > 0, \\ \frac{\Gamma(s_\gamma) - \Gamma(\gamma)}{\Gamma(\gamma)}, \ \textit{if } v_s = 0, \\ \in [0,1] \ \textit{if } v_s = \emptyset. \end{cases}$$

The equilibrium social sanction is $f^* = \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1)$.[21]

Part (i) corresponds to a small κ and/or a large γ. This could be a relatively tolerant society debating a moderately important issue (where the loss of dignity from keeping silent is large relative to the pressure to conform,) or a weak majority facing a relatively large minority as stated in the condition $\gamma \leq min\{\gamma_0, \underline{\gamma}\}$. It could also be a survey conducted by unknown interviewers via telephone, which is a medium with quite small perceived sanctions, a case of small-κ. The equilibrium strategies in (2) are illustrated below; the bold segments represent expressed opinions.

[21]Proofs of propositions 1 and 2 are standard and follow the similar arguments. I prove part (i) here. According to the belief system, only the expressions in $(s_c, 1]$ are sanctioned. Equilibrium payoffs are: All $s \in [0, s_c]$ get zero, individuals $s \in (s_c, \gamma)$ get $-|s - s_c|$ whereas the minorities get $-f^*$. Consider a minority member. Recall that by Definition 1 $s_c = \gamma - f^*$, thus, the minority member will not deviate to any sanction-free majority expression $v < \gamma - s_c$. Deviating to silence avoids the sanction but yields $-\alpha$, which is also not beneficial because by $\gamma \in (\hat{s}, min\{\gamma_0, \underline{\gamma}\}]$, we have $\alpha > f^*$. Finally, choosing another sanctioned minority opinion only decreases his payoff below $-f^*$.

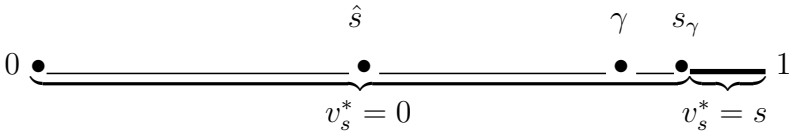
Consider now the majority members $s \in (s_c, \gamma)$ whose own opinions are sanctioned if expressed. According to the equilibrium, these individuals express s_c. The best deviation is to express their own opinion, which yields $-f^* < -|s - s_c|$, hence is not beneficial. Nor is silence a beneficial deviation because $\alpha > f^*$. Finally, all majority members in the range $[0, s_p]$ obtain the maximal utility zero by expressing their own opinions. Therefore, all strategies are optimal given beliefs. Consistence of the common belief system with these strategies is obvious.



Because a positive measure of the minority can beneficially express the majority opinions in the range (s_c, γ), in equilibrium social sanctions target these majority opinions as well. Thus, to avoid sanctions, individuals holding opinions in (s_c, γ) all express the opinion s_c. Those who express the opinions in the range $[0, s_c]$ are correctly interpreted as majority members whereas those who express in the range $[\gamma, 1]$, as minority members and are sanctioned.

The larger the minority group, i.e., the closer γ to the median $\hat{s}$, the smaller the social sanction and the closer is s_c to γ, implying a smaller range of unvoiced opinions (s_c, γ). The essential characteristics of the equilibrium are preserved as long as the minority remains large, i.e., as long as γ does not exceed $min\{\gamma_0, \underline{\gamma}\}$. What happens beyond this upper bound depends on which of the two critical levels, γ_0 and $\underline{\gamma}$, is smaller.

The case $\gamma_0 < \underline{\gamma}$ and $\gamma > \gamma_0$, presented in part (ii) of Proposition 1, corresponds to $\alpha > \gamma_0$. Compared with part (i), we have a large and less tolerant majority, the debate involves a more difficult issue and higher social tensions, but individuals also feel an obligation to express an opinion. This is a combination of large γ and large κ, hence a large social sanction, which nevertheless is smaller than the cost of silence α. Under these circumstances the equilibrium strategies in (3), illustrated below, are played. The entire majority as well as minority members from $[\gamma, s_\gamma]$ comply with the opinion $s = 0$ and all expressions except $s = 0$ are sanctioned. The sacrifice of expressive utility from compliance with $s = 0$ is too large for minority members from $(s_\gamma, 1]$. Given also the large cost of silence, these individuals choose to express their own sanctioned opinions. The resulting highly polarized expression outcome has a wide range of silenced opinions except those in the extremes, where a silent "don't know" answer is rather unacceptable, approximating the national debates on separatist movements in the Russian Federation and Ceylon in the 1990s, or the debate on the Armenian question in Turkey.



For smaller minority sizes, i.e., larger levels of γ, the per-victim social sanction becomes larger, s_γ shifts to the right and a larger population complies with $s = 0$ in equilibrium. Depending on the cost of silence α, further reductions in the range of minority opinions can lead to one of the two possibilities: Either $s_\gamma = 1$ and the population is transformed into a monophonic chorus of $s = 0$, or a new equilibrium emerges in which some individuals opt for silence.

3.2 Equilibria with Silence

Outside the range of parameters admitted in Proposition 1 the social sanction exceeds the cost of silence. The type-1 equilibrium in which all individuals are vocal collapses because majority members from the left neighborhood of γ will deviate to silence, which by Condition B2 they can do without fear of a sanction. The presence of silent majorities drags minority members to silence. Before verifying these observations in Proposition 2, I define as a last step a critical opinion s'_c:

Definition 5 *For $F > \alpha$, let $s'_c = max\{0, \gamma - F\}$.*

Recall, F is a sanction threat, as opposed to the imposed sanction in type-1 equilibria which depends on the relative minority size. The opinion s'_c is the analogue of s_c in Definition 1, but smaller than s_c because the sanction threat F in Definition 5 exceeds the sanction f used in Definition 1. The interpretation of s'_c is then similar: the sanction-free opinion expression which yields the borderline individual the same utility as his own, sanctioned, opinion γ. Proposition 2 describes *type-2* equilibrium outcomes:

Proposition 2 *(i) If $\underline{\gamma}(\kappa, \alpha) \leq \gamma_0(\kappa)$ and $\gamma > \underline{\gamma}(\kappa, \alpha)$, the expressions equilibrium is:*

$$\textit{Strategies: } v_s^* = \begin{cases} s \text{ if } s \leq s'_c, \\ s'_c \text{ if } s \in (s'_c, s'_c + \alpha], \\ \emptyset \text{ if } s > s'_c + \alpha, \end{cases} \tag{4}$$

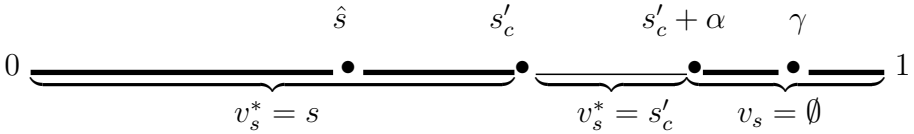
$$\textit{Belief system: } \mu(v_s|\{v^*\}) = \begin{cases} 0 \text{ if } v_s \leq s'_c \\ 1 \text{ if } v_s > s'_c \\ \frac{1-\Gamma(\gamma)}{1-\Gamma(s'_c+\alpha)} \text{ if } v_s = \emptyset. \end{cases}$$

(ii) If $\underline{\gamma}(\kappa, \alpha) > \gamma_0(\kappa)$, $\alpha < 1$ and $\gamma \in (\alpha, 1)$, the expressions equilibrium is:

$$\textit{Strategies: } v_s^* = \begin{cases} 0 \text{ if } s \leq \alpha, \\ \emptyset \text{ if } s > (\alpha, 1], \end{cases} \tag{5}$$

$$\textit{Belief system: } \mu(v_s|\{v^*\}) = \begin{cases} 0 \textit{ if } v_s = 0, \\ 1 \textit{ if } v_s > 0, \\ \frac{1-\Gamma(\gamma)}{1-\Gamma(\alpha)} \textit{ if } v_s = \emptyset. \end{cases}$$

The equilibrium sanction satisfies $F^* > \alpha$.

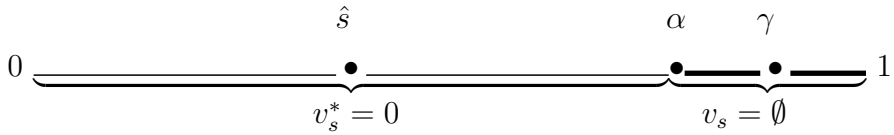


The equilibrium strategies in (4), illustrated above, are associated with large γ and κ, and a relatively small α. These conditions point to combinations of a moderately large majority with a culture that does not tolerate dissenting opinions, imposing and enforcing strict norms on speech. If one takes the example of the cartoon crisis with the publication of Prophet Mohammed's head as a bomb in the Danish press in 2006, the debate involving freedom of the press versus respect for the religions would fit into a type 2(i) equilibrium in a relatively moderate secular Islamic society like Turkey, where in fact the conservative religious views of the majority dominated the media against the few voices for the freedom of expression while many others, presumably intimidated by social sanctions, followed the maxim "when you don't have anything nice to say, don't say anything at all." Simultaneously, the debates on this same issue in several Middle Eastern countries were totally and even violently dominated by the pro-religion opinions of the overwhelming majority and a single extreme pro-religion voice prevailed. That outcome accords with a type 2(ii) equilibrium.

A type 2(i) equilibrium is rich in expression strategies: The majorities from $[0, s'_c]$ do not face a risk of imitation from the minority, so their optimal strategy is to express their own opinions. However, the opinions in the range $(s'_c, s'_c + \alpha]$ can be imitated by the minority (recall, $s'_c = \gamma - F$) and in equilibrium are subject to the sanction F. As a result, individuals holding these opinions express s'_c, which is less costly than silence. The rest, including majority members in the range $(s'_c + \alpha, \gamma)$ as well as the entire minority, are silent. Note that the threat of the sanction F^*, though not imposed on any individual in equilibrium, determines the position of the majority member at $s = s'_c + \alpha$, who is indifferent between silence and expressing

s'_c.[22]

As for the equilibrium in part (ii), it is associated with a larger but nonprohibitive cost of silence, $\alpha \in (\gamma_0, 1)$, a larger κ and/or a smaller minority. This is an environment in which individuals are under the pressure of expressing an opinion on the one hand and a large social sanction on the other hand, thus, conditions are extremely unfavorable to freedom of expression. A wide range of majority members located at the left of α comply with the opinion $s = 0$ whereas the rest of the population is silent. The fraction $[\Gamma(\gamma) - \Gamma(\alpha)]/[1 - \Gamma(\alpha)]$ of the silent group is formed by majority members.



It is also useful to compare Proposition 2(ii) strategies with those of Proposition 1(ii) where a range $[s_\gamma, 1]$ of minority members express their own opinions despite sanctions. The range $[s_\gamma, 1]$ shrinks as γ is raised towards α, and at $\gamma = \alpha$ the borderline minority member becomes indifferent between silence and complying with $s = 0$. For $\gamma > \alpha$, majority members located at the left neighborhood of γ will switch to silence which is sanction-free by Condition B2. This leads to a type-2(ii) equilibrium: majority members from the range (α, γ) plus the entire minority $[\gamma, 1]$ plunge into silence whereas the rest of the majority keeps complying with $s = 0$, supported by a social sanction $F^* > \alpha$ on all expressions except $s = 0$.

The analysis sheds light on the determinants and consequences of social tolerance. It is not controversial to call a society more tolerant than another if under similar parameter configurations the equilibrium range of expressions in the latter is a proper subset of the former.[23] What factors contribute to social tolerance? If

[22] The probability that a silent individual holds a minority opinion is $\mu(\emptyset, \{v\}) = \frac{1-\Gamma(\gamma)}{1-\Gamma(s'_c+\alpha)} < 1$. Hence, silence is not sanctioned because the group includes both minority and majority members.

[23] *Social tolerance* can broadly be defined as "a liberal social attitude towards opinion expressions." Webster's Third International Dictionary defines tolerance as "a permissive or liberal attitude towards beliefs or practices differing from or conflicting with one's own." The larger the number of individuals that are more tolerant in this sense, the more tolerant is the society. Social tolerance is better tested in environments where individuals feel strong incentives to express their opinions, i.e., where the cost of silence is large, for example, under conditions of high issue relevance and awareness. In such environments, expression outcomes are more sensitive to, hence, are more likely to reflect, differences in attitudes towards diversity of opinions.

a tolerant society displays a greater variety of expressions, this can be attributed to a larger minority size factor captured by γ and/or a smaller sanction intensity parameter κ representing other factors contributing to the majority's willingness to punish. Since the social sanction is increasing in both κ and γ, loci of constant sanctions $f(\kappa, \gamma) = c$ can be defined, along which the measure of unvoiced opinions is constant. So, a smaller range of punishable minority expressions by itself does not guarantee a more tolerant outlook−the expressions outcome may well display a smaller variety of expressions. Or, a larger majority with a smaller sanction intensity parameter κ can exercise the same pressure and leave unchanged the measure of absent expressions. Some of the important conclusions of this section are highlighted below.

In expression media with large α and κ (i.e., high issue relevance, face-to-face interactions, punitive norms) individuals conform with rather extreme versions of the dominant majority opinions. Other factors constant, a smaller range of minority opinions (large γ) leads to a larger sanction and further contributes to conformism.

If social tolerance is to be judged by the variety of expressed opinions, large κ and γ are associated with low social tolerance: opinion misrepresentation is common if the cost of silence is large ($\alpha \geq 1$), the minority and its majority neighbors choose silence if the cost of silence is small.

For each reduction in the size of the minority, there is a reduction in the sanction intensity that keeps the measure of silenced opinions unchanged.

4 Dynamics of Opinion Expressions

Up to this point the analysis takes as given a distribution of opinions and explains public expression outcomes. The dynamic extension in this section postulates a link from expression outcomes to distribution of opinions, based on the premise that unvoiced opinions gradually disappear while expressed opinions attract new adherents and grow over time.

Public expressions at odds with private opinions affect the latter. We can more easily rest assured that we see the issue correctly and shall maintain our views when we find social validation in others' expressions. Symmetrically, we are often more likely to change our opinion when we see it disagrees with others' opinions. The mechanism through which individual opinions change under the influence of public expressions works rather in a subconscious manner−a process which Habermas

(1991) terms *social raisonnement*. Models of preference evolution in social settings capture this process often by positing a law of motion, evolving at a speed that depends on the discrepancy between actual behavior and true preferences.[24] Besides the subconscious process of opinion change operating at the individual level, there is another factor that operates over longer periods of time, as younger generations who form their own opinions under the influence of what they hear and see gradually replace the old. The approach adopted below captures these dynamics by imposing week restrictions on the aggregate evolution of opinion distributions.

Let $\Gamma_t(.)$ and $g_t(.)$ denote respectively the cumulative opinion distribution and corresponding density functions at date t. Given the equilibrium expressions $\{v^*\}_t$ at date t, a transition $\Gamma_t \to \Gamma_{t+1}$ generates the opinion distribution at date $t+1$ in accordance with the following property:

(P) *If no individual expresses opinion s at date t, $0 < g_{t+1}(s) < g_t(s)$. If there exists a range of unexpressed opinions and opinion s is expressed at date t, then prob$[g_{t+1}(s) > g_t(s)]$ is positive−equal to one if a positive measure of individuals express s.*

Property (P) merely states that an expressed opinion grows with positive probability and, if a group of individuals all voice the same opinion, then, and only then, the density of that opinion increases with probability one.[25] Nothing is assumed about the magnitude of change in the density of any opinion. But an important implication of (P) is that when a range of opinions is absent whereas some other interval of opinions is expressed, then a subset of that interval will see its density grow in the next round. During this process the individual justifiably expects no perceptible influence from his actual expression strategy to future opinion distributions and outcomes because he is an infinitesimal segment of society. Technically, we have a sequence of expression games where opinion distributions are transformed

[24]The largely subconscious nature of preference evolution is documented in experimental cognitive psychology research. See for example Kahneman and Snell (1992) and Loewenstein and Schkade (1998). Kuran and Sandholm (2008) discuss in some detail the justification and foundations of this approach to evolution of opinions or preferences.

[25]This is a reasonably week assumption to capture the fact that an opinion expressed by only one individual is less likely to attract new adherents than an opinion expressed by many. Note that there can be a countable number of opinions commonly expressed by a (measurable) group of individuals in this model−in fact, in the equilibria displayed in propositions 1 and 2, this number is at most equal to one.

according to property (P). The change in opinions modifies equilibrium expression strategies, which then induce further opinion adaptations.[26]

A final remark. The model parameters (the cost of silence α, sanction intensity κ and the range of minority opinions $[\gamma, 1]$) may change over time and affect the distribution of opinions. For example, persistent falls in the minority population may increase social tolerance by reducing the sanction intensity parameter κ. Property (P) is silent about these potential effects. The analysis is carried out under a fixed configuration of α, κ and γ, which keeps the exposition simple and clear.

Proposition 3 *Under Property (P) a type-1(i) equilibrium remains of type-1(i) at all finite t. As t increases, the social sanction falls, the range of expressed opinions and the minority population grow while a range of majority opinions to the left of γ gradually loses its population.*

A potential transformation of the density of opinions under Property (P) is shown in Figure 2. I provide below the technical details of the proof of Proposition 3; those of propositions 4 and 5 follow along similar arguments. Consider an initial opinion distribution $\Gamma_0(.)$ under which the equilibrium is of type-1(i), where individuals in (s_{c0}, γ) express s_{c0} to avoid the sanction f_0. Under property (P) each opinion in $[0, s_{c0}] \cup [\gamma, 1]$ attracts new adherents with positive probability from (s_{c0}, γ), implying $\Gamma_1(s_{c0}) > \Gamma_0(s_{c0})$ while $\Gamma_1(\gamma) < \Gamma_0(\gamma)$. In the next round the minority grows and the sanction per victim falls:

$$f_1^* = \kappa(\frac{\Gamma_1(\gamma)}{1-\Gamma_1(\gamma)} - 1) < \kappa(\frac{\Gamma_0(\gamma)}{1-\Gamma_0(\gamma)} - 1) = f_0^*.$$

Given the fact that the critical opinion s_c (see Definition 1) is increasing in f, the range of opinions that are not expressed by any individual, now denoted (s_{c1}, γ), shrinks. The date-1 equilibrium remains of type-1(i). The same mechanism will generate at date 2 a smaller range (s_{c2}, γ) relative to (s_{c1}, γ), hence, an increasing sequence of critical opinions $\{s_{c0}, s_{c1}, s_{c2}, \cdots\}$ emerges. In addition, this sequence

[26]The sanction in each period is based on inferences drawn from the expression profile in that period. So, in this large social setting an individual identified in the past as a minority member and sanctioned can, in the present, avoid the sanction by adopting a majoritarian expression strategy. Expectations about future distribution of opinions can affect individuals' willingness to speak out today in isolated small group interactions. See Salmon and Neuwirth (1990) and Scheufele, et al.(2001).

has the property that $s_{c(t+1)} - s_{ct} < s_{ct} - s_{c(t-1)}$ because the growth of the minority population slows down as the range of unvoiced opinions $(s_{ct} - \gamma)$ shrinks.

It is possible to observe an increasingly polarized opinion distribution while the range of expressions narrows down if the sanction intensity parameter κ rises, say, because the issue gains increased relevance for the majority. Proposition 3 points to an alternative evolution where the true opinion distribution becomes bimodal along with a growing variety of public expressions. When the equilibrium is of type-1(i) to start with, the process works to the detriment of unvoiced majority opinions at the neighborhood of the minority group. The minority, then, can grow over time by attracting adherents from this majority group, leading to a fall in the per-victim social sanction. Thanks to the falling sanction, some majority members whose opinions were initially sanctioned may start expressing their own opinions and credibly signal their types. The true opinion density function is transformed by losses of mass from the left of γ in both directions, which leads to a wider range of expressions in conjunction with a relatively polarized opinion distribution.

This model thus shows that social sanctions do not necessarily trigger a process by which the minority vanishes or ends up conforming with the majority. The minority may grow if the initial conditions of type-1(i) equilibria prevail, i.e., in a relatively tolerant society debating a morally loaded issue (relatively large minority, large α and small κ.)

[Figure 2] [Figure 3]

Proposition 4 *An expression environment in which the equilibrium is of type-1(ii) to start with presents a rich class of potential dynamics. The equilibrium is likely to eventually switch to a type-2(ii) equilibrium if $s_{\gamma 0}$ is large, converge (without ever switching) to a type-1(i) equilibrium if $s_{\gamma 0}$ is small, close to γ.*

Recall that in a type-1(ii) equilibrium the population conforms with the expression $s = 0$, except a subset $[s_{\gamma 0}, 1]$ of the minority who express their own opinions. Applied to these initial conditions, Property (P) stipulates increases in the population at $s = 0$ as well as the minority opinions in $[s_{\gamma 0}, 1]$. The critical issue for the dynamics is whether the overall minority population rises or falls. The fate of the minority depends on the position of $s_{\gamma 0}$, which determines the balance between two opposing effects, one that stems from the fact that minority opinions in the range

$[\gamma, s_{\gamma 0})$ are becoming less populated, the other from the fact that those in the range $[s_{\gamma 0}, 1]$ are becoming more populated (see Figure 3.)

If the overall minority population falls ($\Gamma_1(\gamma) < \Gamma_0(\gamma)$,) in the next round the social sanction increases and as a result, $s_{\gamma 1} > s_{\gamma 0}$, leading to a smaller vocal minority group at date $t = 1$. The larger $s_{\gamma 0}$ (the larger the conforming subgroup of the minority), the stronger is the case for a falling minority population and a rising sanction per-victim and, as a result, for $s_{\gamma 1}$ to exceed $s_{\gamma 0}$, $s_{\gamma 2}$ to exceed $s_{\gamma 1}$, and so on. In this case, there are two possibilities: For $\alpha < 1$ the process moves toward a type-2(ii) equilibrium where the entire minority switches to silence along with neighboring majority members, as the social sanction per victim will eventually exceed the cost of silence. For $\alpha \geq 1$ the equilibrium remains of type-1(ii) but displays increased conformity with $s = 0$ over time. On the other hand, if $s_{\gamma 0}$ is small, close to γ, the endogenous sanction is more likely to fall than rise and therefore $s_{\gamma t}$ is likely to keep approaching γ and the process, to move toward the region of type-1(i) equilibria. In this case the initially small conforming segment of the minority reacts to a general loosening of sanctions and gradually switches to expressing true opinions, which leads to an increase in the overall minority population.

The debate on abortion can serve to illustrate the interplay of the factors shaping public opinion distribution as well as the dynamics in Proposition 3. The U.S. opinion distribution on abortion remained strikingly stable over a period of three decades.[27] The median American citizen approves abortion in cases of rape, fatal fetus defect and protection of the mother's health but rejects it otherwise. According to Jelen and Wilcox (2003) we have a slightly bimodal opinion distribution, where a stable majority favors legal abortion at least under some circumstances, a large minority approving abortion without reservation and a relatively smaller categorically opposing group at the other tail. Such a balanced initial opinion distribution can hardly be associated with a large social sanction per victim. On the

[27]See for example Jelen and Wilcox (2003) and the references therein. Though several surveys are available, the most comprehensive in time and scope is the General Social Survey with six questions on opinions about abortion designed to place individual opinions on a scale from the extreme conservative (pro-life) to the extreme liberal (pro-choice) position. Using this data, Jelen and Wilcox (2003) confirm the findings of previous studies about the aggregate stability of the opinion distribution over a period of three decades (1972-2002), with minor ebbs and flows caused by events such as Supreme Court decisions (Roe v. Wade 1973, Webster 1989).

other hand the issue is emotional but technically easy and public awareness is high, which are conditions that favor expressions and raise the cost of silence. Therefore it is not surprising to observe that individuals prefer expressing their opinions without much fear of isolation, leading to a persisting type-1(i) equilibrium rich in expressions with virtually unchanged opinion distribution, in particular a large minority generally unfavorable to abortion preserving its size.

The same-sex marriage issue is another case in point, but with a rapidly changing public opinion. Lax and Phillips (2009) and Gelman, Lax and Phillips (2010) document the dramatic shift, which in the case of New York moved from 36 percent support in 1994-6 to 42.5 percent in 2003-4 and finally to 51 percent in 2008-9, whereas the figures for Utah are respectively 12, 14 and 16 percent and for Illinois, 25, 34 and 41 percent. Nationwide a 25 percent minority supporting same-sex marriage against a majority of which the extremes demand legislation to forbid recognition is transformed within two decades into a public of "two minds," with 45 percent support for same-sex marriage. Though the evolution of expression outcomes has not been documented systematically, the parameters underlying the debate point to a type-1(i) equilibrium: the issue is technically easy (high awareness) and value-laden, corresponding to a large cost of silence α as in the abortion debate. But the relative minority size, affecting the social sanction per victim, is initially smaller in the same-sex marriage debate. Accordingly the present model suggests a considerably smaller sanction today than 15 years ago for signaling a pro-marriage position, in particular in states such as Massachusetts, New York, Maine and California. Transition to a large minority along with diminishing social sanctions is consistent with the dynamics of a vocal type-1(i) equilibrium. The picture in Utah, on the other hand, fits better into a type-1(ii) equilibrium pattern in Proposition 4 (the case of small s_γ) where the large conservative majority keeps a tiny vocal minority under pressure, with a broad range of unvoiced mild opinions losing adherents in both directions.[28]

[28]The U.S. opinion shift is paralleled internationally. Smith's (2011) sample of 31 countries indicates moderate to strong bimodal opinion distributions and an annual two percent average growth in support of same-sex marriage during the last 20 years. In almost all the countries in the sample the second largest group facing a majority take the opposite polar position. Mild polarization dynamics is consistent with those of a type-1 equilibrium. Smith reports substantial differences between the countries as well. In contrast with the 69.6 percent Dutch expression that homosexual behavior is "not wrong at all," the expression outcome in Turkey consists of a 90.8 percent for "always wrong," 2.9 percent for "almost always wrong," 1.8 percent for "wrong only

Consider, finally, type-2 equilibria at the outset. While the opinion adaptation process that initiates from a type-1(ii) equilibrium is path-dependent, the evolution of a type-2 equilibrium is relatively simple and predictable:

Proposition 5 *If the initial equilibrium is of type-2, under Property (P) the majority increasingly dominates the opinion climate: In a type-2(i) equilibrium, majority opinions in* $[0, s_c']$ *will grow, whereas in a type-2(ii) equilibrium the society will eventually convert to the opinion* $s = 0$.

Proposition 5 deals with the case of large κ and a small or moderately large α. Under these conditions, Property (P) implies that a widening tendency to self-censor will exclude minority viewpoints whereas publicly expressed majority opinions will increasingly attract attention and new adherents over time. If the date-0 equilibrium is of type-2(ii) where the only expressed opinion is $s = 0$, in the absence of external shocks or interventions to model parameters κ, α and γ, the society will converge to extreme conformism by building density at $s = 0$. Similarly, if the equilibrium is of type-2(i) to start with, the majority opinion group $[0, s_c']$ will grow to the detriment of unvoiced opinions from the range $(s_c', 1]$. This process will eventually "transfer" the entire society to the left of s_c'.

These dynamics accord with surveys of the American public opinion on school integration. Race-sensitive issues are likely to be of large-κ, involving high social sanctions, so we should expect distortions in expressions even in the relatively isolated medium of survey interviews. The American public opinion was almost evenly split in 1956, but by the early 1980s surveys indicated large majority support for school integration (Hochschild and Scott (1998)). This shift to the strong majority position is in line with Proposition 5. However a type-2 equilibrium dynamics also involves an increasing measure of silent individuals. Berinski (1999, 2002) studied the individual motives behind silence on school integration over the period 1972-92 using National Election Survey (NES) data which contains a question with an explicit "silence option."[29] A surprisingly large 35.3 percent of respondents in 1992

sometimes" and 2.1 percent "not wrong al all," abstention remaining at a low 2.4 percent. The resulting pattern in Turkey is likely to be of type-1(ii), or of type-2 involving a silent group, where an overwhelming conservative majority, coupled with the large κ associated with conservative religious lifestyles, would imply an extremely large social sanction in expression media.

[29]Berinski is interested in the broader question as to whether surveyed individuals opting for silence do so by pressure or lack of information or interest. He focused on the school integration question and side-tested whether similar expression strategies are observed in two other

chose the "don't know" answer whereas in 1972 only 18 percent declined to answer the same question. The silent group includes individuals who truly have no opinion besides those who oppose government intervention and/or integration policy but bow to social pressures in the interview. Berinski finds that the rise in silence largely comes from strategic motives. The dynamics of a type-2(i) equilibrium suggest that the growth in silence over this period is partly due to a widening range of individuals opting for silence, partly due to shifts in public opinion, absent opinions dying to populate the expressed majoritarian opinions. The process feeds, and is fed by, rising social sanctions.

5 Conclusions

This paper aims at understanding the conditions under which individuals who value the right to opinion expression misrepresent their opinions or keep silent, and who these individuals are. It studies a model in which the majority imposes a social sanction on individuals who reveal agreement with a range of unacceptable minority opinions. This per-victim sanction is increasing in the size of the majority. The equilibrium predictions of the model are based on relative minority size, a sanction intensity parameter capturing other factors that affect the social sanction, and the individual cost of silence.

When individuals perceive a large cost from silence relative to the social sanction, in equilibrium minority members express their own sanctioned opinions whereas the majority splits into two groups: those who freely express their opinions and those who shift their expressions toward more orthodox opinions in order to avoid

race-sensitive issues, the 1989 New York city preelection survey data and the fair employment question. The school integration question in NES reads, "Some people say that the government in Washington should see to it that white and black children are allowed to go to the same schools. Others claim that this is not the government's business. Have you been concerned enough about this question to favor one side over the other?" [If yes,] "Do you think the government in Washington should see to it that white and black children go to the same schools, or stay out of this area, as it is not their business?" Berinski notes that "don't know" responses are high relative to other racial policy questions and uses statistical techniques that account for selection bias to attribute the difference to the presence of a full filter which explicitly provides an escape from a substantive answer. He also points out that the filter gives "respondents who are uncomfortable expressing anti-egalitarian sentiment an easy way to avoid answering the question: rather than put themselves in a socially difficult position, the respondent can pass on answering the question altogether."

imitation by the minorities. If the sanctioned opinion range becomes smaller the per-victim sanction becomes larger, leading some minority members to also start conforming with the extreme orthodox majority.

Silence arises as an equilibrium strategy when individuals do not perceive an *involvement obligation*, for example, in expression media such as the internet or survey interviews where the cost of silence is relatively small. Typically the silent group is composed of the minority plus neighboring majority members. Keeping the cost of silence small, the model predicts that a wider range of majority opinions are silenced in debates involving a smaller range of sanctionable minority opinions.

The dynamic extension of the model generates a rich class of predictions about possible evolutions of the public opinion distribution, based on the assumption that absent opinions gradually lose density to expressed opinions. Accordingly, if the social debate opens with an equilibrium involving a vocal minority, this minority should grow over time while the social sanction falls, leading to an increasingly tolerant climate of expressions despite possibly an increasingly polarized public opinion. If the debate opens with an equilibrium involving silent individuals, the process should lead to conformity with a subset of majority opinions, which shrinks to a singleton−the extreme orthodox majority opinion−if the social sanction is very large to start with.

This continuum-agent model applies to large social settings where individuals cannot expect to influence social expression outcomes. Small-group debates can be quite different and richer in expression strategies such as social loafing, i.e., to keep silent based on the expectation that others will express similar opinions, an option which would reduce individuals' willingness to speak out. In repeated small-group interactions, individuals can expect to affect the climate and opinion distributions, including their own future opinion and ability to express themselves truthfully, which interestingly complicates formulation of expression strategies. Extensions to the static game are also open: allowing for additional sources of population heterogeneity would enrich the model and its equilibrium expression outcomes. For instance, differential costs of silence would introduce the possibility for two individuals with identical opinions to adopt different expression strategies: one individual expressing his own opinion, the other choosing conformity or silence.[30]

[30]The assumption of homogeneous preferences is likely to overestimate the majority's ability to shut down extreme minority opinions. This is so because extremists may have a relatively strong preference for expressing their opinions and so may also be less vulnerable to social sanctions

The analysis also excludes the possibility of a positive net social sanction exported by a powerful minority group.[31] Finally, it is worth noting that in this model speech itself does not have a silencing effect. As noted by Fiss (1996), speech may have a silencing effect if it reduces the resources available to others to pass their voices.[32] Competition for speech-enabling resources may call for state intervention, a controversial issue which remains to be explored in broader models.

A large literature, sampled in the references, specializes in developing and experimentally testing hypotheses on opinion expressions. This paper's contribution is partially to offer a unifying formal framework with refined predictions about how the individual and environmental parameters it identifies affect static expression outcomes and their evolution. I close the paper with short remarks about experimentally testing these predictions.[33] First, to accurately measure distortions to expressions the participants must be offered a rich set of expression choices, including silence. It is not uncommon that studies including the silence option add only a binary agree-disagree choice, or omit silence when several expression choices are availed. The second issue, also noted by Berinski (1999), is that because social pressures and lack of awareness are both conducive to silence, the design should minimize the compositional bias from "don't know" responses. Supplementary questions can be formulated to disentangle the motive behind silence or, as some recent experiments do, survey methods to elicit individual and group characteristics can be combined with small chat room experiments to directly observe expression strategies without imposing a strict response format. An example is Parker (2009) who finds that vocal individuals are predominantly at the two extremes and those who suppress their opinions are located at the middle range of the opinion distribution, which is consistent with type-1 equilibria in Proposition 1. Third, a proper test requires reliable measures of the cost of silence relative to the cost of opinion misrepresentation. Extant empirical studies attempt to estimate either one or the

(Noelle-Neumann (1993).)

[31] See Centola et al (2005) for small-group enforced norms.

[32] Fiss draws a difference between a street-corner speaker and officially financed exhibition of artistic work: while the former does not seem to be consuming a public resource for expression, the latter is.

[33] Real-world data on national debates can be more problematic because they are plagued by interventions and shocks to the medium of expression. Nevertheless, whether expressions evolve in accordance with Property (P) can be tested indirectly from the news media and opinion leaders, with data on who is covered by the press while speaking up and who is not, combined with regular surveys to measure the true opinion distribution and other model parameters.

other cost but not both. When a range of opinions is absent, the relative magnitude of the two costs determines the type of the equilibrium, that is, whether individuals in that range are silent or misrepresent their opinions.

Appendix

A. Equilibria with pooling in expressions

In this section I expose the structure of equilibria in which a group of minority and a group of majority express the same opinion s_p. This type of equilibria can be constructed when $\alpha \geq f$, the case where silence is dominated as in the vocal equilibria stated in Proposition 1. To illustrate, consider the parameter range admitted in Proposition 1(i), i.e., $\gamma \in (\hat{s}, min\{\gamma_0, \underline{\gamma}\}]$. For any such γ, pick a minority opinion $s_m > \gamma$ and a majority opinion $s_p \in (s_c, \gamma)$ such that $s_m - s_p = \gamma - s_c$, and define $\mu_p = [\Gamma(s_m) - \Gamma(\gamma)]/[\Gamma(s_m) - \Gamma(s_p)]$. The social sanction is as stated in Proposition 1, $f^* = \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1)$, and the following is an expressions equilibrium:

$$\text{Strategies:} \quad v_s^* = \begin{cases} s, \text{ if } s \leq s_p \text{ or } s \geq s_m, \\ s_p, \text{ if } s \in (s_p, s_m); \end{cases}$$

$$\text{Belief system:} \quad \mu(v_s|\{v^*\}) = \begin{cases} 1, \text{ if } v_s > s_p, \\ \mu_p \text{ if } v_s = s_p \\ 0, \text{ if } v_s < s_p, \\ \in [0,1] \text{ if } v_s = \emptyset \end{cases}$$

According to these beliefs, all expressions in $(s_p, 1]$ are sanctioned whereas those in $[0, s_p]$ are not. To verify the equilibrium, recall that by Definition 1 $s_c = \gamma - f^*$, thus, $s_m - s_p = f^*$. All minorities located at $s \geq s_m$ express their own sanctioned opinions and obtain the utility $-f^*$. Clearly, they will not deviate to another sanctioned expression in the range $(s_p, 1]$. Since $\gamma \in (\hat{s}, min\{\gamma_0, \underline{\gamma}\}]$, we have $\alpha > f^*$, so none of these sanctioned individuals will deviate to silence. Nor will they deviate to sanction-free expressions in the range $[0, s_p]$ because $-|v - s'| < -f^*$ for any $v \in [0, s_p)$ and $s' \in [s_m, 1]$. As for individuals $s \in [s_p, s_m)$, expression of s_p is optimal and yields the utility $-|s - s_p|$ which exceeds the utility $-f^*$ from expressing own opinion as well as silence. Finally, all majority members in the range $[0, s_p]$ obtain the maximal utility zero by expressing their own opinions. Therefore

the strategies are optimal given beliefs. Consistence of the common belief system with these strategies is obvious.

This equilibrium differs from Proposition 1(i) in one respect. The interval of unvoiced opinions (s_c, γ) shifts to the right and becomes (s_p, s_m) but its size remains the same, f^*.

B. Properties of expressions equilibria

I begin with a result on the equilibrium belief system, followed by monotonicity of expression strategies. A useful observation is that the equilibrium utility of a vocal individual s is bounded below by the sanction: $U_s^* \geq -f$. Any individual can guarantee this utility by expressing his own opinion.

Claim 1. *Fix an equilibrium in which v_A and v_B are two expressed opinions such that $v_B < v_A$. If $\mu(v_B|.) = 1$, then $\mu(v_A|.) = 1$.*

Proof. Suppose, contrary to the claim, $\mu(v_A|.) < 1$, implying that v_A does not trigger a sanction, i.e., $\iota(v_A) = 0$. Clearly, then, in equilibrium the individual located at v_A must be expressing his own opinion. Denote this individual by s. On the other hand, denote the individual who expresses the sanctioned opinion v_B by s'. Since $\mu(v_B|.) = 1$, this individual must be a minority member and, moreover, optimality of his expression strategy implies that he is located at $s' = v_B$, which yields the payoff $-f$. Thus, $v_{s'}^* = s' = v_B \geq \gamma$. Since $v_B < v_A$ by assumption, it follows that $\gamma < v_A$. Because the equilibrium strategy of s' must be utility maximizing, $-f \geq -|v_A - s'|$.

To show that $\mu(v_A|.) = 1$, consider any majority member $t < \gamma$ and let U_t^* denote his equilibrium payoff. Since $t < \gamma \leq v_B = s' < v_A$, we have

$$U_t^* \geq -f \geq -|v_A - s'| > -|v_A - t|.$$

Therefore in equilibrium no majority member expresses v_A. Since v_A is expressed, it must be expressed by a minority member, implying $\mu(v_A|.) = 1$, a contradiction.

The next result establishes monotonicity of equilibrium expression strategies.

Claim 2. (monotonicity) *Consider two individuals s, s' such that $s' > s$. If in equilibrium $v_s^* \in [0,1]$ and $v_{s'}^* \in [0,1]$, then $v_{s'}^* \geq v_s^*$.*

Proof. Assume, on the contrary, $v_{s'}^* < v_s^*$. Optimality of v_s^* and $v_{s'}^*$ imply, respectively,

$$|v_s^* - s| + \iota(v_s^*)f \leq |v_{s'}^* - s| + \iota(v_{s'}^*)f, \quad \text{and} \tag{6}$$

$$|v_{s'}^* - s'| + \iota(v_{s'}^*)f \leq |v_s^* - s'| + \iota(v_s^*)f. \tag{7}$$

By Claim 1, the case $\{\iota(v^*_{s'}) = 1, \iota(v^*_s) = 0\}$ is ruled out. Consider the case $\iota(v^*_{s'}) = \iota(v^*_s)$, i.e., both expressions are sanctioned or neither is. From Condition (6) we get $v^*_{s'} < s$ (if $v^*_{s'} \geq s$, we have $v^*_s - s > v^*_{s'} - s \geq 0$ since $v^*_s > v^*_{s'}$, which violates (6).) From Condition (7), in turn, we get $v^*_s > s'$ (for if $v^*_s \leq s'$, then, $v^*_{s'} - s' < v^*_s - s' \leq 0$ because $v^*_s > v^*_{s'}$, hence $|v^*_{s'} - s'| > |v^*_s - s'|$, which violates (7).) Therefore, $v^*_{s'} < s < s' < v^*_s$. Using this fact, (6) and (7) can be written as:

$$|s - s'| + |s' - v_s| \leq |v_{s'} - s| \text{ and } |v_{s'} - s| + |s - s'| \leq |v_s - s'|.$$

Using the first inequality in the second leads to $|s - s'| \leq -|s - s'|$, which is impossible.

The last possibility is $\{\iota(v^*_{s'}) = 0, \iota(v^*_s) = 1\}$; v^*_s is sanctioned whereas $v^*_{s'}$ is not. The individual s who expresses a sanctioned opinion must be expressing his own opinion, i.e., it must be that $v^*_s = s$. Thus, $v^*_{s'} < v^*_s = s < s'$. The optimality conditions, analogues of (6) and (7), can be written as:

$$f \leq |v^*_{s'} - s| \quad \text{and} \quad |v^*_{s'} - s'| \leq f.$$

These conditions imply $s \geq s'$, a contradiction. Thus, $v^*_{s'} \geq v^*_s$.

Equilibrium expression strategies have other properties which I explore below. The following definition deals with a particular expression outcome.

Definition. *A* pool *under a strategy profile* $\{v\}$ *is a closed interval* $[z, z']$ *and an opinion* $y \in [z, z']$ *such that* $v_x = y$ *if and only if* $x \in [z, z']$..

That is, a pool $\{[z, z'], y\}$ consists of a group of individuals who all express the same opinion $y \in [z, z']$ and no outsider expresses an opinion from the interval $[z, z']$. A pool is called a *right pool* if $y = z'$, a *left pool* if $y = z$ and a *centered pool* if $y \in (z, z')$.

Claim 3. *In any equilibrium, a pool must be a left pool.*

Proof. To show a contradiction, let there be an equilibrium $(\{v^*\}, \mu^*, f)$ with a right of centered pool $[z, z']$. As a preliminary observation, first, in such equilibria we must have $\mu(y|\{v^*\}) < 1$, hence, $\iota(y) = 0$. For if $\mu(y|\{v^*\}) = 1$, any individual $x \in [z, z']$ would deviate to $v_x = x$ and increase his utility by $|x - y|$. Second, we must have $\mu(x|\{v^*\}) = 1$ for all $x \in [z, z']$ and $x \neq y$, for otherwise expression of the own opinion x would yield a higher payoff.

There are three cases according to whether the pool involves only minority members, only majority members, or a mixture of both. I consider these cases in order below.

(*All-minority pool*) If $\{[z, z'], y\}$ is a pool and $z \geq \gamma$, the pool includes no majority members and, necessarily, $\mu(y|\{v^*\}) = 1$. Clearly, $U_x^* < -f$ for all $x \neq y$ and $x \in [z, z']$. Then, however, $v_x^* = y$ is not optimal, contradicting the assumption that v_x^* is an equilibrium strategy.

(*All-majority pool*) Suppose $[z, z'] \subseteq [1, \gamma)$. Let U_γ^* denote the equilibrium payoff of individual γ and recall, U_γ^* is bounded above by zero. Define $\tilde{s}_c = max\{0, \gamma + U_\gamma^*\}$. The individual at γ obtains the same utility as his equilibrium utility U_γ^* by expressing $\tilde{s}_c$, provided $\iota(\tilde{s}_c) = 0$ ($\tilde{s}_c$ is not sanctioned).

If $\tilde{s}_c > z$, by Condition B1 we have $\mu(z|\{v^*\}) < 1$, so the individual at z can obtain the maximal utility zero by deviating to $v_z = z$. If $\tilde{s}_c \leq z$, we have $-|\gamma - \tilde{s}_c| = U_\gamma^* < -|\gamma - y|$, which implies that the individual at γ earns a higher payoff from $v_\gamma = y$, which is not his equilibrium strategy, a contradiction.

(*Mixed pool*) Consider now a pool containing both majority and minority members, with $z < \gamma \leq z'$. There are two cases, $y < \gamma$ and $y \geq \gamma$.

(a) Suppose $y < \gamma$. Then, $-|\gamma - x| < U_\gamma^* = -|\gamma - y|$ for any $x < y$. The equilibrium payoff of any minority member, including the individual at γ, is larger than the payoff from expressing x smaller than y. For any $x \in [z, y)$, however, we have $-|y - z| < -|y - x|$, hence, individual z can beneficially deviate to expressing x if x is not sanctioned. Indeed, because no minority member would benefit from this deviation, it follows by Condition B1 that $\mu(x|\{v^*\}) < 1$. The individual z will therefore deviate to $v_z = x \in [z, y)$, contradicting optimality of $v_z^* = y$.

(b) Suppose $\gamma \leq y$ and define $\xi = |\gamma - y|$. The arguments in Part (a) above imply that if $\gamma - \xi > z$, then $v_z^* = y$ cannot be optimal. I consider the other two cases below:

(i) Suppose $z \geq \gamma - \xi > 0$. An opinion $x \in [\gamma - \xi, z)$ at the left neighborhood of the pool is either expressed in equilibrium, or it is not.

Suppose x is expressed by some individual k. We know that k cannot be a minority member, so $k < \gamma$, because all minority members prefer expressing y to expressing k outside the pool. Since x is expressed and those who express it are not minority members, x cannot be sanctioned: $\mu(x|\{v^*\}) < 1$ and $\iota(x) = 0$. If x is not sanctioned, in equilibrium it must be expressed by the individual located at x (i.e., $k = x$.) Then, however, the individual z in the pool will deviate to $v_z = x$ and obtain the payoff $-|z - x| > -|x - y|$, which upsets the equilibrium.

Suppose x is not expressed in equilibrium. In this case, because it is a majority

opinion, expression of x must be sanctioned in equilibrium: $\mu(x|\{v^*\}) = 1$. Consider now the left neighborhood of $\gamma-\xi$. Since $z < \gamma$, there exists $k \in (max\{0, z-\xi\}, \gamma-\xi)$ such that $-|y-z| < -|z-k|$. Therefore, individual z from the pool will deviate to k if $\iota(k) = 0$. Indeed, since $|\gamma-k| < U^*_\gamma$ and $-|z-k| > U^*_z = -|y-z|$, Condition B1 implies $\mu(k|\{v^*\}) < 1$, hence, $\iota(k) = 0$. It follows that individual z will deviate to $v_z = k$, which upsets the equilibrium by contradicting optimality of $v^*_z = y$.

(ii) The last case is $\gamma - \xi \leq 0$. Because $\iota(0) = 0$ by assumption (expression of the extreme majority opinion is never sanctioned by the majority,) the individual at z from the pool will deviate to $v_z = 0$: we have $-|\gamma-y| \leq -|\gamma-0|$ which implies $-|z-y| < -|z-0|$ hence upsets the equilibrium.

So far I established that if equilibrium expression strategies generate a pool, this must be a left-centered pool $\{[z, z'], z\}$, where $y = z$. Next, I show that if in equilibrium an individual located at s expresses an opinion $z < s$, then, individuals located between z and s also express z.

Claim 4. *$v^*_s = z < s \Rightarrow v^*_x = z$ for all $x \in [z, s)$.*

Proof. Note that the belief system must satisfy $\mu(z|\{v^*\}) < 1$, for otherwise the equilibrium payoff of individual s would be below the lower bound $-f^*$, a contradiction. Also, $\mu(x|\{v^*\}) = 1$ for any $x \in (z, s]$, because otherwise the individual s would prefer deviating to $x \in (z, s]$. Given these beliefs, I claim that the individual located at any such x cannot be silent in equilibrium. If he were, $U^*_x = -\alpha \geq -|x-z|$ and because $s > x$, it follows that $-\alpha > -|s-z|$, which contradicts the fact that $v^*_s = z$ is optimal. Given that any individual located at x expresses an opinion, by monotonicity we get $v^*_x \leq z$, and since z is not sanctioned, expression of z must be optimal for an individual located at $x \in [z, s)$.

Next, I show that if in equilibrium two distinct individuals s and s' both choose silence, so do all individuals located between s and s'.

Claim 5. *Suppose, in equilibrium, there exist individuals $s, s' \in [0, 1]$ such that $s < s'$ and $v^*_s = v^*_{s'} = \emptyset$. Then, $v^*_y = \emptyset$ for all $y \in (s, s')$.*

Proof. Assume, contrary to the claim, that there exists some $y \in (s, s')$ such that $v^*_y \neq \emptyset$. Then, necessarily, $v^*_y \in (s, s')$. To see this, suppose that $v^*_y \leq s$. It follows that $-|y-v^*_y| - \iota(v^*_y)f > -\alpha - \iota(\emptyset)F$. But then $v^*_s = \emptyset$ cannot be optimal because $-|s-v^*_y| - \iota(v^*_y)f > -\alpha - \iota(\emptyset)F$, a contradiction. A similar argument implies that $v^*_{s'} = \emptyset$ cannot be optimal if $v^*_y \geq s'$.

Now choose s and s' such that $[s, s']$ is the smallest interval containing in its

strict interior the individuals who express an opinion. By monotonicity and the fact that the strict interior of $[s, s']$ cannot contain a right pool or a centered pool, all individuals in this interval must be expressing their own opinions, i.e., $v_x^* = x$ for all $x \in (s, s')$. Consider the individual $z = s + \epsilon$ with ϵ arbitrarily small. We know that $v_z^* = z$. If $\iota(z) = 0$, then $U_z^* = 0$ but $v_s^* = \emptyset$ cannot be optimal. If $\iota(z) = 1$, then, either $v_s^* = \emptyset$ or $v_z^* = z$ is not optimal, because $\alpha + \iota(\emptyset)F > f$ and $\alpha + \iota(\emptyset)F \leq f$ cannot be both be true. In either case, a contradiction is established.

Using the result in Claim 5, I show below that any group of silent individuals must lie at the right end of the interval $[0, 1]$, thus, including minority members.

Claim 6. *$v_s^* = \emptyset \Rightarrow v_{s'}^* = \emptyset$ for all $s' > s$.*

Proof. Suppose there exist $s' > s$ such that $v_{s'}^* \in [0, 1]$ and let z be the largest k such that $v_k^* = \emptyset$. We thus have $v_x^* \in [0, 1]$ for all $x > z$ and, by Claim 5, $z < s'$.

Consider the individual at $x = z + \epsilon$ where $\epsilon > 0$ is sufficiently small. By monotonicity and Claim 3 (no right or centered pool exists) $v_{z+\epsilon}^* = z + \epsilon$. This expression must be sanctioned, for if it is not, the individual at z would deviate to expressing $z + \epsilon$ and increase his payoff from $-\alpha - \iota(\emptyset)f$ to $-\epsilon$, which is arbitrarily close to the maximal payoff zero. If the expression $z + \epsilon$ is sanctioned, by optimality of individual strategies we have $-\alpha - \iota(\emptyset)f > -f$ for the individual z (because he can guarantee the payoff $-f$ by expressing his own opinion,) and $-f \geq -\alpha - \iota(\emptyset)f$ for the individual $z + \epsilon$. Thus, either $v_{z+\epsilon}^*$ or v_z^* is not optimal.

The nest result will complete the characterization of equilibria.

Claim 7. *In any equilibrium there exists at most one left pool.*

Proof. Suppose there are two left pools $[z_1, z_1']$ and $[z_2, z_2']$ such that $z_1' < z_2$. Recall that left pools have the property that $\iota(z_1) = \iota(z_2) = 0$ and $\iota(x) = 1$ for all other x in the two pools, which imply the equilibrium payoffs $U_{z_1}^* = U_{z_2}^* = 0$.

By Claim 6, individuals located between the two pools cannot be silent. By the assumption that there is no pool between the two pools, all individuals in (z_1', z_2) must be expressing their own opinions. Consider an individual between the two pools, $s \in (z_1', z_2)$ arbitrarily close to z_2. The expression s should not be sanctioned for otherwise the individual s would deviate to expressing z_2 and obtain the payoff $-|z_2 - s|$, which is impossible because he does not belong to this pool. By the same logic, individuals at the left neighborhood of s cannot be sanctioned either. Applying this logic successively implies that all expressions $x \in (z_1', z_2)$ must be sanction-free. Then, however, z_1' can increase his utility from $-|z_1' - z_1| < 0$ to $-\epsilon$

by deviating to the expression $z_1' + \epsilon$, which contradicts optimality of $v^*_{z_1'} = z_1$.

The results so far establish that equilibrium expression strategies satisfy monotonicity, that there can be at most one left pool (where an interval of individuals all express the opinion at the lower end of the interval) and that in equilibrium the set of silent individuals, if any, is an interval of the form $[z, 1]$. Therefore the only possible equilibrium type when the cost of silence exceeds the social sanction ($\alpha \geq f$, see Definition 3) must be of type-1, stated in Proposition 1, where all individuals express an opinion. When $\alpha < f$, however, Condition B2 rules out vocal type-1 equilibria because there exist majority members who prefer silence over their expressions. Thus, for these parameter values all equilibria are of type-2 and involve silence.

References

[1] Akerlof, G, A. (1980) "A Theory of Social Custom, of which Unemployment may be one Consequence," *Quarterly Journal of Economics* 94: 749-775.

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

[3] Berinsky, A.J., (1999)"The Two Faces of Public Opinion" *American Journal of Political Science* 43: 1209-1230

[4] Berinsky, A.J., (2002) "Political Context and the Survey Response: The Dynamics of Racial Policy Opinion," *The Journal of Politics* 64: 567-584

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

[6] Black, D. (1948). "On the Rationale of Group Decision-making," *Journal of Political Economy* 56: 23-34.

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

[8] Cho, I-K. and Kreps, D. (1987) "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics* 102: 179-221.

[9] Coleman, S., (2004) "The Effect of Social Conformity on Collective Voting behavior," *Political Analysis* 12: 76-96.

[10] Dharmapala, D. and McAdams, R. H., (2005) "Words that Kill? An Economic Model of the Influence of Speech on Behavior (with Particular Reference to Hate Speech)," *Journal of Legal Studies* 34: 93-136.

[11] Fiss, O. M., (1996) *The Irony of Free Speech* Harvard University Press: Cambridge, Massachusetts

[12] Gelman, A., Lax, G. and Phillips, J. (2010) "Over time, a gay marriage groundswell" *New York Times* August 22. 2

[13] Gerber, A.S., Green, D.P, and Larimer C.W. (2008) "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment," *American Political Science Review* 102: 33-48.

[14] Glaeser, E. L., (2005) "The Political Economy of Hatred," *Quarterly Journal of Economics* 120: 45-86.

[15] Habermas, J. (1991) *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society.* Cambridge, MA, MIT Press.

[16] Harrison, T. (1940). "What is public opinion?" *The Political Quarterly*, 11: 368-383.

[17] Hayes, A. F., Shanahan, J. and Glynn, C. J., (2000). "Willingness to Express One's Opinion in a Realistic Situation as a Function of Perceived Support for that Opinion," *International Journal of Public Opinion Research*, 13: 45-57.

[18] Hayes, A. F., Glynn, C. J., and Shanahan, J. (2005). "Willingness to self-censor: A construct and measurement tool for public opinion research." *International Journal of Public Opinion Research*, 17: 298-323.

[19] Ho, S. S, and McLeod, M. D. (2008) "Social-Psychological Influences on Opinion Expression in Face-to-Face and Computer-Mediated Communication," *Communication Research* 35: 190-207.

[20] Hochschild, J. L and Scott, B. (1998) "Governance and Reform of Public Education in the United States. *Public Opinion Quarterly* 62:79-120.

[21] Jelen, T.G. and Wilcox, C. (2003) "Causes and Consequences of Public Attitudes Toward Abortion: A Review and Research Agenda," *Political Research Quarterly* 56: 489-500.

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

[23] Kahneman, D. and Snell, J. (1992), "Predicting a Change in Taste: Do People Know What They Will Like?", *Journal of Behavioral Decision Making*, 5: 187-200.

[24] Kuran, T. (1987) "Preference Falsification, Policy Continuity and Collective Conservatism," *The Economic Journal* 97: 642-665.

[25] Kuran, T. (1995) *Private Truths, Public Lies.. The Political Economy of Preference Falsification*, Chicago, IL, University of Chicago Press.

[26] Kuran, T. and Sandholm, W.H. (2008) "Cultural Integration and Its Discontents," *The Review of Economic Studies* 75: 201-228.

[27] Lapinski, M. K., and Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, 15, 127-147.

[28] Lax, J.R., and Phillips, J.H. (2009) "Gay Rights in the States: Public Opinion and Policy Responsiveness" *American Political Science Review* 103, 367-386.

[29] Lazear, E., P. (1999) "Culture and Language," *Journal of Political Economy* **107**, 95-126.

[30] Loewenstein, G. and Schkade, D. (1998), "Wouldn't It Be Nice? Predicting Future Feelings", in D. Kahneman, E. Diener and N. Schwarz (eds.) *Well Being: The Foundations of Hedonic Psychology*, New York: Russell Sage Foundation, 85-105.

[31] Madison, J. (1961) "The Federalist No 49, February 9, 1788," pp. 338-47 in Jacob E. Cooke Ed., *The Federalist.* Middletown, Conn.: Wesleyan University Press.

[32] McDevitt, M, Kiousis, S, and Wahl-Jorgensen, K (2003) "Spiral of Moderation: Opinion Expression in Computer-Mediated Discussion," *International Journal of Public Opinion Research* 15: 454-470.

[33] Noelle-Neumann, E. (1974) "The Spiral of Silence: A Theory of Public Opinion," *Journal of Communication* 24: 43-51.

[34] Noelle-Neumann, E. (1979) "Public Opinion and the Classical Tradition; A Re-evaluation," *The Public Opinion Quarterly* 43: 143-156.

[35] Noelle-Neumann, E. (1993) *The Spiral of Silence: Public Opinion—or Social Skin*, Chicago, IL, University of Chicago Press.

[36] Parker, D. J. (2009) "Avoiding Groupthink: Whereas Weakly Identified Members Remain Silent, Strongly Identified Members Dissent About Collective Problems," *Psychological Science* 20: 546-548.

[37] Salmon, C. T., and Neuwirth, K. (1990). "Perceptions of opinion climates and willingness to discuss the issue of abortion". *Journalism Quarterly* 67: 567-577.

[38] Scheufele, D. A. and Eveland, W.P Jr. (1999) "Perceptions of 'Public Opinion' and 'Public' Opinion Expression," *International Journal of Public Opinion Research* 13: 25-44.

[39] Scheufele, D. A. and Moy, P. (2000) "Twenty-Five Years of the Spiral of Silence: A Conceptual Review and Empirical Outlook," *International Journal of Public Opinion Research* 12: 3-28.

[40] Scheufele, D. A., Shanahan, J., and Lee, E. (2001). "Real talk: Manipulating the dependent variable in the spiral of silence research" *Communication Research* 28: 304-324.