Opinion Expressions under Social Sanctions

Mehmet Bac*

March 24, 2009

Abstract

I study a social debate where individuals are subject to informal sanctions if their expressions 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 shut opinion expressions, leave unchecked the extreme variants of the dominant majority and, potentially, homogenize expressions. While diversity of expressions is generally praised for conducing to 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 extreme crimes such as rape 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] These are discrete choice models with heterogeneous agents whose individual payoffs increase when others behave similarly. Often individual choice sets include only two options, which may be appropriate for the study of phenomena such as elections, segregation or political participation where individual choice sets are naturally discrete and small. The context of opinion expression differs in that each individual can express from a wide range of opinions using different words and tones, so the choice set is (ap-

[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] Over the past few decades a number of papers have extended the critical mass models 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) and Chwe (1999). Brock and Durlauf (2001) develop a generalized critical mass model with microfoundations.

proximately) a continuum, to which one must add the option of remaining silent. 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] 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 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—the smaller the target group relative to the majority, the larger is the harm per victim. 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

[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 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. Zaller (1992) develops a theory to explain how people receive political information and determine their opinions.

[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.

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.

To highlight some of the model's predictions, in equilibrium sanctions can be ineffective on a large minority. Interestingly, 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. A small dose of informal sanctions immediately builds a gap between the expressions of the two camps by inducing the majority neighbors to distance themselves from the minority. In general, the set of sanctioned expressions is never confined to the minority range−majority opinions that come sufficiently close to the minority are also sanctioned.

The key determinant of the identities of silent and vocal individuals is size of the social sanction relative to the cost of silence. 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 not voiced in public. If the social sanction becomes large yet remains smaller than the cost of silence, equilibrium expressions may even display a greater variety of minority opinions than majority opinions. When the social sanction rises, the expression gap between the two groups widens 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 that are located at the opposite extreme of the minority. 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.

Silence becomes an equilibrium strategy if the minority is small, provided, of course, the cost of silence is not too large. I impose a "right to silence" condition on beliefs about the types of silent individuals, which allows majority members to become silent if they wish so, without fear of a social sanction. Under this condition, a silent group always consists of the entire minority plus a range of majority neighbors. Moreover, in equilibrium silent individuals are never sanctioned (if silence were sanctioned, it would be dominated by truthfully expressing own

opinions, which is a costless option whereas silence entails a psychic cost).

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 conform with the majority. Minority opinions are likely to keep loosing their adherents if the initial equilibrium involves a silent minority, which the model is associated with a small group of minority opinions in a public debate over a morally loaded issue (a cost of silence that is large, yet smaller than the social sanction.)

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 privately known, but the distribution of opinions is public knowledge.[5] It is also public knowledge that an opinion $\gamma \in (\hat{s}, 1)$ is the borderline separating the society in two camps. Opinions in the range $[\gamma, 1]$ are in minority−these could be views sympathizing with faith-based violence, favoring a totalitarian political regime or expansion of a cultural minority's political rights. The analysis admits all $\gamma > \hat{s}$ and delivers a predicted outcome of expressions for any range of minority opinions.

The individual *expression strategy*, $v_s : [0,1] \to [0,1] \times \emptyset$, assigns to each $s \in [0,1]$

[5]Public knowledge of the distribution simplifies the analysis. Introducing prior beliefs about opinion distributions would substantially and unnecessarily complicate the analysis.

an opinion from $[0,1]$ or silence, $\{\emptyset\}$. Then an *expression outcome* is a profile of expression strategies $\{v\}$, 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 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

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

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

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

The expressive utility of individual s is single-peaked at s and silence generates the disutility $\alpha > 0$ [6]

A social sanction is imposed on individuals who are commonly perceived to hold a minority opinion. Given a profile of expressions $\{v\}$, a common belief system assigns to each expression v_s a probability $\mu(v_s|\{v\})$ that speaker s holds a minority opinion; that is, $\mu(v_s|\{v\}) = prob(s \geq \gamma|\{v\})$. I assume that individual s is sanctioned only if he reveals a minority type with probability one, that is, $\iota(v_s) = 1$ only if $\mu(v_s|\{v\}) = 1$.[7]

As for the size of the per-victim social sanction, I adopt a simple functional form. This specification distinguishes between two cases: the case where the sanction is

[6] $|x|$ 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, $-(v_s - s)^2$. 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. I comment on the impact of heterogeneous costs of silence in the concluding section.

[7] The qualitative results continue to hold if social sanctions are imposed for beliefs $\mu \in [\overline{\mu}, 1]$ where $0 < \overline{\mu} < 1$. See Bernheim (1994) for more general belief systems.

actually imposed on a commonly identified minority member and the case where it is not.

$$f = \begin{cases} \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1), \text{ if there exists } t \geq \gamma \text{ such that } \mu(v_t|.) = 1, \\ F > 0, \text{ otherwise.} \end{cases} \quad (1)$$

When minorities are correctly identified, relative group size matters and the sanction takes the form in the first line of (1).[8] $f = \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1)$ is a convenient form; alternative specifications are admissible. The intensity parameter $\kappa > 0$ represents determinants of the sanction other than the relative group size effect. In case no individual is identified as minority and sanctioned, the second line of (1) applies: then the sanction F is a threat on individuals who deviate to a range of "off-the-equilibrium," unvoiced, opinions. Determination of F and the expression range in which the sanction applies 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 $v^*_s \in [0, 1]$, then $|v^*_s - s| < \alpha$.

Conditions B1 and B2 overcome the problem of multiplicity of equilibria by

[8]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.

imposing reasonably plausible restrictions on beliefs concerning *off-the-equilibrium* expression strategies. They also 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). It concerns equilibria involving vocal minority members: 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 with probability one. The belief system should assign $\mu(t|\{v^*\}) < 1$, i.e., expression of the opinion t should be sanction-free. This condition isolates equilibria in which a majority member is artificially induced to keep silent or to express a different opinion despite the fact that his opinion cannot beneficially be mimicked by a minority member.

Condition B2 can be termed majority members' "right to silence." It imposes a restriction on beliefs about the types of silent individuals. It would be unnatural, if not impossible for a majority to punish silent individuals when its own members prefer remaining silent over their actual expression strategies. B2 eliminates equilibria in which the population is forced to speak out by fear of an artificial sanction on silence. On the other hand, the right to silence is not a safe heaven for the minority; silent or vocal, the latter is subject to sanctions whenever correctly identified. But 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.

Finally, I set aside equilibria in which all individuals are silent[9] and I adopt a tie-breaking convention: If an individual is indifferent between silence and expressing an opinion $v_s \in [0, 1]$, then he expresses v_s. Indifference between expressing own sanctioned opinion and a different but sanction-free opinion is broken in favor of expressing own opinion.

[9] A simple way to guarantee that at least one opinion will be expressed in this model is to assume that the farthest opinion from the punishable minority range, $v = 0$, is sanction-free. To illustrate why it is not reasonable to have a sanction on $v = 0$, this would be tantamount to assuming that the author of an extreme anti-racist statement is ostracized by the anti-racist majority, because from this statement people infer that the author is racist with probability one.

3 Expressions Equilibria

The first part of this section presents some definitions and elementary results. Throughout the analysis the borderline individual $s = \gamma$ plays an important role. This is reflected in the definitions of the two critical opinions, s_c and s_s (see Figure 1.)

Definition 1

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

The borderline individual is indifferent between expressing the sanction-free majority opinion s_c and his own sanctioned opinion, γ. He is also indifferent between silence and expressing the majority opinion s_s. A negative s_s therefore indicates that silence is strictly dominated (then, expressing the extreme opposite opinion $s = 0$ yields a larger payoff to all minority members.) Since $s_s > 0$ if and only if $\alpha < \gamma$, it follows that s_s is positive for all $\gamma \in (\hat{s}, 1)$ provided $\alpha < \hat{s}$.

[Figure 1]

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.*

Any factor that increases the per-victim social sanction induces the borderline individual to accept a larger sacrifice of expressive utility. The impact of κ on s_c is therefore expected. That of γ may not be so. If the sanction f were fixed in size, γ and $s_c = \gamma - f$ would always move in the same direction. But according to (1) the sanction is a function of the relative size of the minority, so, 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. This total effect is unambiguously negative for γ sufficiently large. When the range of minority opinions shrinks, the per-victim sanction becomes very large, which reduces s_c and eventually makes it negative.

The critical opinion γ_0 defines a borderline individual who is indifferent between expressing $s_c = 0$ and expressing his own (sanctioned) opinion:

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

As γ approaches the median opinion $\hat{s}$ from above, the size of the minority approaches that of the majority, as a result the social sanction per victim vanishes and s_c also approaches $\hat{s}$.

The position of s_s relative to s_c is an important analytical consideration. The following definition is useful in this regard.

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

When s_c and s_s are both positive and the range of minority opinions is given by the interval $[\underline{\gamma}(\kappa, \alpha), 1]$, the borderline individual $\underline{\gamma}(\kappa, \alpha)$ becomes indifferent between three options: silence, expressing his own sanctioned opinion, and expressing the sanction-free majority opinion s_c. Note that $\underline{\gamma}(\kappa, \alpha)$ is well-defined.[10] Lemma 2 highlights some important properties of $\underline{\gamma}(\kappa, \alpha)$ and $\gamma_0(\kappa)$.

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

(ii) $s_s < s_c$ if and only if $\gamma < \underline{\gamma}(\kappa, \alpha)$.

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

Part (ii) concerns the case of a large minority, hence a small per-victim sanction. In this case the borderline individual γ prefers expressing the majority opinion s_c over silence, if both options are sanction-free. By part (i), if $\alpha \leq \gamma_0(\kappa)$, i.e., if the cost of silence is not large, there exists a critical borderline individual $\gamma = \underline{\gamma}(\kappa, \alpha)$ who is indifferent between three options, expressing his own opinion, the majority opinion $s_c > 0$ and remaining silent. This case is illustrated in Figure 1. On the other hand, if $\alpha > \gamma_0(\kappa)$, the borderline individual's best alternative to expressing his own sanctioned opinion is to express the sanction-free majority opinion s_c provided $s_c > 0$, that is, provided $\gamma \leq \gamma_0(\kappa)$. Remaining in the case $\alpha > \gamma_0(\kappa)$, an increase in γ (reducing the range of minority opinions) will raise the social sanction and eventually lead s_c to zero. Finally, part (iii) states the impact of the sanction intensity parameter κ on the two critical γ values. It is worth noting that a higher κ reduces s_c but leaves s_s unchanged, so, makes silence more attractive as an option to escape from the social sanction.

[10] 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.

When silence is a strictly dominated option ($\alpha \in (\gamma_0(\kappa), 1]$) despite a large per-victim social sanction (or a small minority, $\gamma > \gamma_0(\kappa)$), there exists a critical minority opinion s_γ, defined below, separating the population in two groups: those who prefer expressing their own sanctioned opinions to the sanction-free opinion $v = 0$ and those who prefer the opposite.

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

Note that in this definition a subset of the group who prefers expressing the sanction-free opinion $v_s = 0$ to expressing their own sanctioned opinions consist of minority members from the range $[\gamma, s_\gamma)$.

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 $\kappa = 0$.[11] All equilibria involve a mixture of separation and pooling in either silence or expression of a specific opinion.

In *type-1* equilibria, every individual voices an opinion. The equilibria in part (i) of Proposition 1 are partially separating in the sense that all minorities voice their own opinions, are correctly identified and sanctioned. In part (ii) equilibria a positive measure of minority members, along with the majority, comply with the expression $v = 0$.

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

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

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

[11] 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.

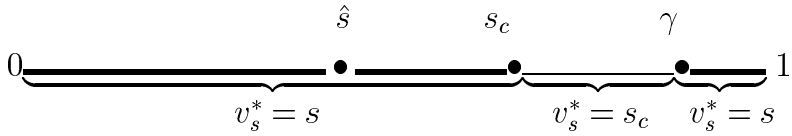
(ii) If $\gamma_0(\kappa) < \underline{\gamma}(\kappa, \alpha)$ *and* $\gamma \in (\gamma_0(\kappa), min\{\alpha, 1\})$*, the expressions equilibrium is:*

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

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

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

Type-1 equilibria arise when the cost of silence exceeds the social sanction, which is a likely case under a small sanction intensity parameter κ and/or a large minority. Equilibrium strategies are given by (2) when the minority is "sufficiently" large, more precisely, when γ does not exceed $min\{\gamma_0(\kappa), \underline{\gamma}(\kappa, \alpha)\}$. In this case, s_c is nonnegative and not smaller than s_s. The opinions expressed in these equilibria are illustrated below by the bold segments.

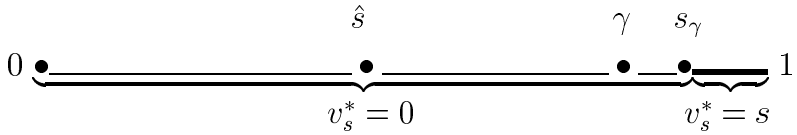


Note that the social sanction is not confined to minority opinions; each majority opinion in the range (s_c, γ) can beneficially be mimicked by a positive measure of the minority and so, is also subject to sanctions if expressed. To avoid this, all individuals holding opinions in the range (s_c, γ) all express the opinion s_c. Those who express the opinions in the range $[0, s_c]$ are correctly interpreted as majority members while those who express in the range $[\gamma, 1]$, as minority members and are sanctioned. The strategies in (2) are optimal given the belief system and the belief system is consistent with strategies and satisfies conditions B1 and B2−no majority member is forced to speak by a sanction threat on silence.

While almost all opinions are voiced in equilibrium when γ is close to $\hat{s}$, beyond the point at which s_c is maximal, further increases in γ quickly widens the range of self-censored opinions, as shown in Figure 2 by the vertical distance between the γ line and the s_c schedule. The essential characteristics of the equilibrium are preserved as long as γ does not exceed $min\{\gamma_0(\kappa), \underline{\gamma}(\kappa, \alpha)\}$. What happens beyond

this upper bound depends on which of the two critical levels, $\gamma_0(\kappa)$ and $\underline{\gamma}(\kappa, \alpha)$, is smaller.

The case $\underline{\gamma}(\kappa, \alpha) \leq \gamma_0(\kappa)$ and $\gamma > \underline{\gamma}(\kappa, \alpha)$ is treated in Proposition 2(i). As for the case $\gamma_0(\kappa) < \underline{\gamma}(\kappa, \alpha)$ and $\gamma > \gamma_0(\kappa)$ presented in part (ii) of Proposition 1, it corresponds to $\alpha > \gamma_0(\kappa)$, i.e., a negative s_c (see Lemma 2.) The cost of silence is very large but the range of minority opinions is very small, which implies a very large social sanction per victim. Under these circumstances the equilibrium strategies in (3), illustrated below, are played. All opinion expressions except $s = 0$ are sanctioned. The entire majority as well as minority members from $[\gamma, s_\gamma]$ comply with the opinion $s = 0$. 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.



For smaller minority sizes, i.e., larger levels of γ, the per-victim social sanction becomes very large, s_γ shifts to the right and a larger population complies with $s = 0$ in equilibrium. A point will be reached where 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. Which of the two outcomes prevails depends primarily on α.

3.2 Equilibria with Silence

Outside the range of parameters admitted in Proposition 1 the social sanction exceeds the cost of silence. Then, the type-1 equilibrium collapses because majority members from the left neighborhood of γ will prefer to deviate from expressing the sanction-free opinion, $s = 0$ or s_c, to silence, which for these deviants must also be sanction-free by Condition B2. The new equilibrium must involve silent majority members and, in addition, a sanction must be pending on a range of expressions to make silence an optimal strategy for some individuals. When these conditions obtain and a positive measure of majority members choose silence, the entire minority must also be silent in equilibrium. Before verifying these observations in Proposition 2, I define as a last step a critical opinion s_c':

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

Note that s'_c is the analogue of s_c in Definition 1, but smaller than s_c because the sanction F in Definition 5 exceeds the sanction 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:*

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

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

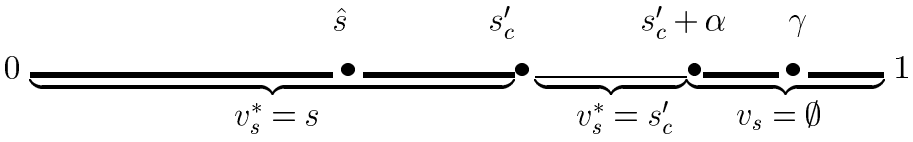
(ii) If $\underline{\gamma}(\kappa, \alpha) > \gamma_0(\kappa)$, $\alpha < 1$ and $\gamma \in (\alpha, 1)$, the expressions equilibrium is:

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

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

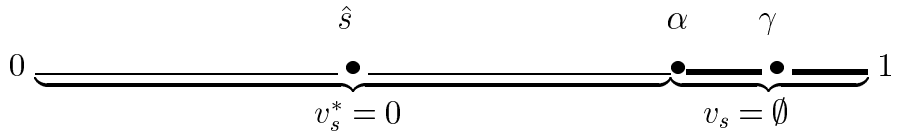
The equilibrium sanction satisfies $F^* > \alpha$.

In type-2 equilibria, there is a silent group consisting of the entire minority plus a fraction of the majority. A sanction that exceeds the cost of silence is pending on a range of expressions, but no individual is actually sanctioned. Among the individuals whose opinions are subject to sanctions, some keep silent while the rest choose to express the sanction-free opinion s'_c.



The equilibrium strategies in (4), illustrated above, are associated with high γ, small α and relatively large values of κ. The equilibrium is rich in expression strategies: Majority members from the range $[0, s'_c]$ face neither a risk of imitation from the minority nor sanctions if they express their own opinions; their optimal strategy is therefore to express their own opinions. However, the opinions in the range $(s'_c, s'_c + \alpha]$ can be imitated by the minority and, so, are subject to the sanction F^*. Fearing the sanction, individuals holding these opinions adopt the pooling strategy $v_s^* = s'_c$. The remaining group, including majority members in the range $(s'_c + \alpha, \gamma)$ as well as the entire minority $[\gamma, 1]$, chooses silence. 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.[12]

Finally, the equilibrium in part (ii) is associated with a large but nonprohibitive cost of silence ($\alpha \in (\gamma_0(\kappa), 1)$,) a large κ and/or a small minority. This is an environment in which individuals are under pressure to express an opinion and a large sanction is imposed on identified minority members, thus, conditions are extremely unfavorable to freedom of expression and, to a lesser extent, keeping silent. The parameter configurations are similar to those associated with type-1(i) equilibria, except that the cost of silence is smaller than the endogenous sanction. However, this leads to a large difference in equilibrium strategies: majority members located at the left of α continue to comply with the opinion $s = 0$ while the rest switch to silence. The fraction $[\Gamma(\gamma) - \Gamma(\alpha)]/[1 - \Gamma(\alpha)]$ of the silent group is formed by majority members.



One can gain further insight into the conditions leading to this equilibrium by reconsidering Proposition 1(ii) strategies, the case $\gamma \in (\gamma_0(\kappa), \alpha)$ assuming $\alpha < 1$. There, the range of minority members who express their own opinions despite sanctions, $[s_\gamma, 1]$, shrinks as γ is raised towards α, and at $\gamma = \alpha$ the borderline minority member becomes indifferent between complying with $s = 0$ and silence. For $\gamma > \alpha$, the right to silence (condition B2) of majority members located at the

[12]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.

left neighborhood of γ upsets the type-1 equilibrium.[13] The new equilibrium is of type-2(ii): 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 expression of any opinion except $s = 0$.

Existence of equilibrium follows from the fact that the two propositions cover the entire parameter space. In the Appendix I show that the equilibrium is unique up to the sanction F imposed on off-the-equilibrium expressions in type-2 equilibria.[14]

4 Synthesis and Implications

An alternative and useful classification of equilibria is obtained by varying the cost of silence relative to other parameters. The case $\alpha \leq \gamma_0(\kappa)$ (cost of silence is small relative to sanction intensity) is covered by propositions 1(i) and 2(i): In the former, we have a relatively large minority (γ in the range $(\hat{s}, \underline{\gamma}(\kappa, \alpha)]$) and a range of opinions represented by the vertical segment between the γ line and the s_c curve in Figure 2 are not voiced. Type-2(i) equilibria arise when $\gamma > \underline{\gamma}(\kappa, \alpha)$, in the case of small minority, hence, large social sanction: A group $(s_c' + \alpha, 1]$ of individuals choose silence while the set of expressed opinions shrinks to $[0, s_c']$.

The case $\alpha > \gamma_0(\kappa)$, equivalently, $\underline{\gamma}(\kappa, \alpha) > \gamma_0(\kappa)$ (cost of silence is large relative to sanction intensity) corresponds to propositions 1(ii) and 2(ii). In the former, we have a large minority and the entire population conforms with the expression $v = 0$ except a subset of the minority located at the other extreme. In the latter, we have $\alpha \in (\gamma_0(\kappa), 1)$, i.e., a large but not prohibitive cost of silence coupled with a small minority ($\gamma > \alpha$). The sanction, therefore, is very large and the entire minority along with majority members from the range (α, γ) are silent, whereas majority members whose opinions are in the range $[0, \alpha]$ express $v = 0$.

[Figure 2] [Figure 3]

[13] The majority member located at $\gamma - \epsilon$ gets $-|\gamma - \epsilon|$ by expressing $s = 0$, $-\alpha$ by remaining silent. Since $\gamma > \alpha$, silence is obviously a better option for ϵ small enough.

[14] The qualification to uniqueness bears solely on Proposition 2 equilibria with silence. The magnitude of the sanction F^* on expressions at the right tail of the opinion distribution is part of equilibrium data and determines the range of majority members in $[0, s_c']$, where $s_c' = \gamma - F^*$, who express their own opinions. In contrast with the sanction $f = \kappa(\frac{\Gamma(\gamma)}{1-\Gamma(\gamma)} - 1)$ in vocal type-1 equilibria, no restriction is imposed on F^*. Given an $F^* > \alpha$, equilibrium expression strategies are unique.

In the remainder of this section I relate the model's data to concepts and variables which an interdisciplinary body of research identifies as relevant and operational. I also associate combinations of α, κ and γ with social tolerance and attributes of media of expression. 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 right to expression, α can be linked to at least three factors: issue relevance, issue awareness and the medium of expression.[15]

(i) *Issue relevance.* The cost of silence is large in contexts in which individuals generally feel a strong incentive to express an opinion, for example, in a highly controversial, morally loaded discussion.[16]

(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. Confidence in opinions strengthens the incentives to express and raises the cost of silence. 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* is affected by 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, in mediated internet chatrooms involvement obligation is low thanks to reduced social presence and lack of contact; anonymity of expressions reduces the cost of silence and offers the participants greater latitude to express extreme opinions.[17]

[15]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.

[16]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.)

[17]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.

• **The sanction intensity parameter κ.** The intensity of the social sanction depends on the majority's ability and willingness to punish. The willingness to punish, in turn, is affected by social and economic factors such as openness and a permissive culture towards unorthodox expressions, the likelihood of involvement in potentially beneficial exchange with other citizens and the benefits from these interactions. Expect a small κ in open societies. Expect also a small κ in societies that institute stronger networks for the citizens to develop and deepen social and economic interdependence.[18] 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 parameter. As α, κ should be positively associated with issue relevance and issue awareness. While these two parameters of course do not always move together (for example, a strong and widely shared culture that gives a respectful hearing to those who do not share the same views has a small κ but a relatively large α), it is useful to explore the implications of a positive correlation between κ and α in this model.

Consider, then, a context in which both α and κ become large, holding γ fixed−imagine that the range of punishable minority opinions remains unchanged while the issue becomes more controversial and individuals feel a stronger involvement obligation. A larger κ pivots the s_c schedule in Figures 1-3 clockwise, as a result, shifts both $\gamma_0(\kappa)$ and $\underline{\gamma}(\alpha, \kappa)$ to the left, toward $\hat{s}$. The increase in α, in turn, pulls down the s_s schedule and shifts $\underline{\gamma}(\alpha, \kappa)$ to the right. Subject to two opposing forces from κ and α, the combined effect on $\underline{\gamma}(\alpha, \kappa)$ is ambiguous, though presumably small in magnitude−at least compared with the fall in $\gamma_0(\kappa)$. The final result therefore depends on the initial equilibrium type and on the absolute as well as

It is worth noting that in computer-mediated discussions individuals experience less of the fear or anxiety they do in direct, face-to-face, communication. While this effect enhances the incentives to voice an opinion, hence, raises the cost of remaining silent in internet-mediated discussions, Ho and McLeod's (2008) results suggest a more important role for the involvement obligation effect which works in the opposite direction. The fear or anxiety from direct communication with others is named "communication apprehension" by McCroskey (1977).

[18] 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.

relative changes in α and κ. If both parameters are increased by a small amount in an equilibrium of type-1(i) to start with, we get a wider range of censored opinions because a larger measure of majority members shall pool at expressing a smaller opinion s_c. If κ and α increase substantially, the equilibrium switches to type-1(ii) where minority members from the lower tail of $[\gamma, 1]$ start conforming with the extreme majority opinion at $s = 0$. Or, if the society experiences a large increase in κ combined with a small increase in α, the entire minority may switch to silence−a type-2 equilibrium.

• **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. Given two societies with identical κ, α and opinion distribution $\Gamma(.)$, it is possible to observe the high-γ society in a type-2 equilibrium and the low-γ society in a type-1 equilibrium if the two γ values are sufficiently apart (because per-victim social sanctions could be large enough to silence the minority opinions in the high-γ society.)

It is not controversial to call a society more tolerant than another if under similar conditions of issue relevance and awareness, the equilibrium range of expressions in the latter is a proper subset of the former.[19] What factors contribute to social tolerance? If a tolerant society displays a greater variety of expressions, it should also display smaller per-victim sanctions because a larger social sanction can only reduce

[19] *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.

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	$+$	$--$	$-$

the equilibrium variety of expressions. Using the fact that the endogenous social sanction in (1) is increasing in both κ and γ, loci of constant sanctions $f(\kappa, \gamma) = c$ can be defined, where $c = \gamma - s_c$, along which the measure of unvoiced opinions is constant. Thus, if the level of tolerance is to be judged by the variety of observed expressions, tolerance is constant along an $f(\kappa, \gamma)$ locus and, in a type-1 equilibrium to start with, to any reduction in the sanction intensity parameter one can associate a smaller minority size that keeps constant the measure $\gamma - s_c$ of unvoiced opinions. The important conclusion is, reducing the list of punishable expressions does not in general make a more tolerant society—the expressions outcome may well display a larger social sanction per victim and a smaller variety of expressions. Then a fall in the sanction intensity parameter is necessary to generate an increased variety of expressions.

Table 1 summarizes the expected qualitative relations between characteristics of expression outcomes and the parameters of the present model. Some of the important conclusions are as follows:

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

Large κ and γ are associated with low social tolerance: the minority and its majority neighbors will choose silence unless silence is prohibitively costly (unless $\alpha \geq 1$).

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

5 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.

The claim that public expressions affect private opinions is not controversial. We can more easily rest assured that we see the issue correctly and shall be more inclined to 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 strongly disagrees with others' opinions. The mechanisms through which individual opinions are influenced by public expressions−a process which Habermas (1991) terms *social raisonnement*−are beyond the objective of this paper. I confine this section to determination of possible evolutions of expressions equilibria, assuming a reasonable process governing alterations of opinions.

Let $\Gamma_t(.)$ and $g_t(.)$ denote respectively the cumulative opinion distribution and corresponding density functions at date t. Given the equilibrium expressions $\{v_s^*\}_t$ 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) states that an expressed opinion grows with positive probability. If a group of individuals voice the same opinion, then, and only then, the density of that opinion increases with probability one.[20] An obvious and important implication of (P) is that in the presence of a range of unexpressed opinions, if all the opinions in an interval are expressed, then a subset of that interval will see its density grow. Thus, the present dynamic extension consists of a sequence of

[20]This is a reasonable assumption to capture the fact that an opinion expressed by 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.

static expression games where opinion distributions are transformed according to property (P).[21]

A final remark. Changes in the model parameters (the cost of silence α, sanction intensity κ and the range of minority opinions $[\gamma, 1]$) may, over time, affect the distribution of opinions. An effect in the opposite direction is also possible−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.

Propositions 3-5 describe potential dynamics of public expressions and private opinions as a function of the initial conditions and values of α, κ and γ.[22]

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 population of the minority 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 4(i). 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 opinions in the range (s_{c0}, γ), which implies $\Gamma_1(\gamma) < \Gamma_0(\gamma)$. Growth of the minority reduces the sanction per victim and, 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, (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 $\{s_{c0}, s_{c1}, s_{c2}, \cdots\}$. In addition, this sequence has the property that $s_{c(t+1)} -$

[21] The sanction in each period is based on inferences drawn from the expression profile in that period. So, in accordance with the motivating assumption that individual opinions may change over time, 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. The analysis does not capture the potential effect of the expected distribution of opinions in the future on individuals' willingness to speak out today (on this, see Salmon and Neuwirth (1990) and Scheufele, et al.(2001)).

[22] The results are applications of Property (P) to generate sequences of static expressions equilibria. I omit the formal proofs but sketch the arguments in the text.

$s_{ct} < s_{ct} - s_{c(t-1)}$ because, growth of the minority population slows down as the range of unvoiced opinions shrinks, which translates into the new equilibrium as a smaller reduction in the social sanction, hence, a smaller rise in s_{ct} at larger t.

It is possible to observe an increasingly polarized opinion distribution in conjunction with a diminishing range of public expressions if the sanction intensity parameter κ rises over time, which could be because, say, increased issue relevance. However, Proposition 3 suggests the opposite possibility, that the true opinion distribution can become more polarized despite a growing variety of public expressions. When, as in the present exercise, parameters such as κ are constant and the equilibrium is of type-1(i) to start with, Property (P) implies a rising minority population over time and a falling per-victim social sanction. Thanks to the falling sanction, some majority members whose opinions were initially sanctioned may later start expressing their own opinions and credibly signal that they hold majority views. As the range of expressed opinions widens over time, the true opinion density function is transformed by losses of mass from an interval to the left of γ to other opinions, which suggests an increased polarization in the opinion distribution.

Social sanctions do not necessarily trigger a process by which the minority ends up conforming with the majority. This model shows that the opposite is possible and plausible; the minority may grow by attracting its neighbors despite sanctions. This happens if the initial conditions of type-1(i) equilibria prevail, for example, using the interpretations discussed in the previous section, in a tolerant society debating a morally loaded issue (large α and small κ.)

[Figure 4]

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, to converge towards (without ever switching to) type-1(i) equilibria if $s_{\gamma 0}$ is 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 sanctioned opinions. Applied to these initial conditions, Property (P) stipulates increases in the population at $s = 0$ as well as the minority opinions located in $[s_{\gamma 0}, 1]$. If the overall minority population falls ($\Gamma_1(\gamma) < \Gamma_0(\gamma)$,) the social sanction increases and,

as a result, in equilibrium $s_{\gamma 1} > s_{\gamma 0}$, i.e., there will be a smaller vocal minority group at date $t = 1$. Whether the overall minority population will fall depends on the position of $s_{\gamma 0}$ because the direction of the process is determined by the balance between two opposing effects that depend on $s_{\gamma 0}$, 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 4(ii).) The larger $s_{\gamma 0}$, 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 the case of a falling minority population, two possibilities arise: For $\alpha < 1$, the process moves towards a type-2(ii) equilibrium where the entire minority switches to 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 close to γ, the endogenous sanction is more likely to fall than rise, $s_{\gamma t}$ is likely to keep approaching γ and the process, to move toward the region of type-1(i) equilibria. In this case, the initially silent minority group react to a general loosening of sanctions and start expressing their true opinions. Gradually, the minority population increases.

Consider, finally, type-2 equilibria at the outset. While the opinion re-distribution process that initiates from a type-1(ii) equilibrium is path-dependent, the evolution of 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 a large κ and a small or moderately large α. To exemplify, consider a small group of minority opinions in a public debate over a morally loaded issue. 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 it is a type-2(i) equilibrium 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'. Similarly, if the date-0 equilibrium is of type-2(ii) where the only expressed opinion is $s = 0$, the society will converge to extreme conformism by building density at $s = 0$.

6 Conclusions

This paper attempts to understand the conditions under which informal social sanctions can have powerful effects on expressions in public, why individuals who value the right to opinion expression keep silent or express opinions that differ from their own, and who these individuals are. It develops a model in which the majority imposes a social sanction on individuals who reveal their agreement with a range of unacceptable minority opinions. This per-victim sanction is negatively related to the size of the minority group.

The model predicts that when individuals perceive a large cost from silence relative to the social sanction—a likely case if the sanction intensity is small and the minority is large—minority members express their own sanctioned opinions. The majority is split in two groups: those who freely express their opinions and those who shift their expressions toward more orthodox opinions in order to avoid imitation by the minorities. Increasing the social sanction while keeping the cost of silence large, some minority members may also start conforming with the majority. As for the equilibria involving silence, the size of the silent group (the minority plus a group of majority members) depends on the social sanction; actually this sanction is not imposed in equilibrium but stands as a threat on a range of opinion expressions. In other words, silence becomes a neutral yet costly strategy to escape from large social sanctions.

Combinations of small minority, large sanction intensity and small cost of silence are generally unfavorable to expression of minority opinions in public. Narrowing the range definition of the minority given a fixed opinion distribution, or reducing the population holding a fixed range of minority opinions, leads to an increase in the social sanction, so, is also detrimental to opinion expression. Another finding is that the social sanction and the measure of unvoiced opinions remain constant under appropriate increases in the minority population and the sanction intensity parameter. To illustrate, an increase in the number of individuals holding opinions that are subject to sanctions coupled with an appropriately increased dose of morality in the debate will keep unchanged the measure of absent opinions. A large minority in a morally loaded debate, or a relatively unimportant issue coupled with a small minority—according to this model, in both cases a minority member who reveals his type should expect roughly the same social sanction.

The dynamic extension of the model generates a rich class of predictions about

possible evolutions of the distribution of opinions. The extension is based on the observation that individuals are influenced by publicly expressed opinions and may alter their own opinion for one of these. Accordingly, if the social debate opens with an equilibrium involving a vocal minority, the population of this minority should increase in conjunction with a falling social sanction. This generates an increasingly tolerant climate of expressions despite possibly an increasingly polarized distribution of true opinions. However, if the debate opens with an equilibrium involving silent individuals, the process will lead the distribution of opinions toward full conformity with a subset of majority opinions. This subset shrinks to a singleton—the extreme orthodox majority opinion—if the social sanction is very large to start with. Section 4 offers interpretations of the model parameters and their links to key concepts in communication and public opinion research.

I close the paper with extensions and limitations of the analysis. The assumptions on preferences (disutility from silence and expressing a different opinion than one's own) capture individual values attributed to the freedom of expression. As such, the model leaves out the possibility of social loafing, i.e., abstention or remaining silent based on the expectation that others will express similar opinions. Inclusion of a preference for social loafing would reduce individuals' willingness to speak out. Related to this point is the assumption of population homogeneity regarding the cost of silence and the form of expressive utility. Allowing for additional sources of population heterogeneity would enrich the model and its equilibrium expression outcomes. For instance, it would introduce the possibility for two individuals with identical opinions but different sanction disutilities to adopt different expression strategies: one individual expressing his own opinion, the other choosing conformity or silence. Differential costs of silence can also induce differential expressive behavior.[23] Another simplifying assumption of the model is that sanctions are imposed on individuals who are commonly perceived to belong to the minority with probability one. This rules out the possibility that a majority member is mistakenly perceived as minority and is sanctioned. While allowing for small mistakes of identity would not dramatically change my results, it can be relevant in

[23]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 (Noelle-Neumann (1993).) Research on public opinion (e.g., Horner et al. (1998)) showed that the degree of extremism in opinions is correlated with the individual's cost of remaining silent.

other settings designed to study a different set of issues. In one of these, Lagunoff (2001) has shown that the majority group prefers not to impose stringent standards in punishing minorities by fear of mistakenly punishing its own members. Considerations of this kind can reduce the intensity and incidence of social sanctions. The analysis also excludes the possibility of a positive net social sanction exported by a powerful minority group.[24]

Finally, it is worth noting that in this model speech does not have a silencing effect. There is no need to compete for speech-enabling resources. As noted by Fiss (1996), speech may have a silencing effect if it reduces the resources available to others to pass their voices.[25] This kind of competition may call for state intervention, but the regulation of speech by the state is a controversial issue. The conditions for optimal state intervention to foster full and open debates remains to be explored in broader models.

Appendix

This Appendix states and prove several 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, $\mu(v_B|.) = 1$ implies that expression of the opinion v_B is sanctioned, i.e., $\iota(v_B) = 1$. Denote the individual who expresses v_B by s'. This individual must be a minority member and, moreover, be located at $s' = v_B$, obtaining the utility $-f$. If he is located elsewhere and expresses v_B in equilibrium, his payoff would be smaller than $-f$, contradicting optimality of expression strategies. Thus,

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

[25] 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.

$v^*_{s'} = s' = v_B \geq \gamma$. Since $v_B < v_A$ by assumption, it follows that $\gamma < v_A$. Because the equilibrium strategy of s' must be utility maximizing, $-f \geq -|v_A - s'|$.

Consider any majority member $t < \gamma$ with equilibrium utility denoted U^*_t. Since $t < \gamma \leq v_B = s' < v_A$, we have

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

It follows that in equilibrium no majority member expresses the opinion v_A. Since v_A is expressed, it must be expressed by a majority member, implying $\mu(v_A|.) = 1$, a contradiction.

The next result establishes monotonicity of equilibrium expression strategies.

Claim 2. (monotonicity) *If $s' > s$ and in equilibrium $v^*_s \in [0,1]$, $v^*_{s'} \in [0,1]$, 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 $\iota(v^*_{s'}) = \iota(v^*_s)$. Condition (6) implies $v^*_{s'} < s$ (if $v^*_{s'} \geq s$, we have $v^*_s - s > v^*_{s'} - s \geq 0$ since $v^*_s > v^*_{s'}$, which violates (6).) Condition (7), in turn, implies $v^*_s > s'$ (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, $v^*_{s'}$ is not. Since s expresses a sanctioned opinion, he must be expressing his own opinion, i.e., the expression strategy of s must be $v^*_s = s$. 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.

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

Definition. *A* pool *under strategy profile* $\{v\}$ *is a closed interval* $[z, z']$ *with* $y \in [z, z']$ *such that* $v_x = y$ *for* all $x \in [z, z']$ *and* $v_k \notin [z, z']$ *for all* $k \notin [z, z']$.

That is, a pool $\{[z, z'], y\}$ consists of a group of individuals who all express the same opinion $y \in [z, z']$, while 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. *An equilibrium strategy profile* does not *induce a centered pool or a right pool.*

Proof. Fix an equilibrium $(\{v^*\}, \mu^*, f)$ and let $[z, z']$ be a right or centered pool under strategy profile $\{v^*\}$, thus, with $y \in (z, z']$ such that $v_x^* = y$ for all $x \in [z, z']$. As a first step, note that these equilibrium strategies imply $\mu(y|\{v^*\}) < 1$, 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 $|x - y|$. The same argument implies $\mu(x|\{v^*\}) = 1$ for all $x \in [z, z']$.

I consider all three potential pool types according to involvement of minority and/or majority opinions in order−an all-minority pool, an all-majority pool, or a mixed pool including both minority and majority members.

(*All-minority pool*) If $\{[z, z'], y\}$ is a pool and $z \geq \gamma$, 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 $[z, z'] \subseteq [1, \gamma)$. Define $\tilde{s}_c = max\{0, \gamma + U_\gamma^*\}$ (note that U_γ^* is bounded above by zero). 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, $\mu(z|\{v^*\}) < 1$ and the individual at z can obtain the maximal utility zero by deviating to $v_z = z$. If $\tilde{s}_c \leq z$, $-|\gamma - \tilde{s}_c| \leq U_\gamma^* < -|\gamma - y|$, which implies that the individual at γ can increase his utility by deviating to $v_\gamma = y$, a contradiction with the assumption that we have an equilibrium.

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

(a) Suppose $y < \gamma$. Then, for any $x < y$, $-|\gamma - x| < U_\gamma^* = -|\gamma - y|$: the equilibrium payoff of the individual at γ is larger than the payoff he would get by deviating to an x smaller than y in the pool. The same is true for other minority members

in the pool. But for individual z we have $-|y-z| < -|y-x|$ for $x \in [z, y)$, hence, individual z can beneficially deviate to the opinion x. Since 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 $v_z = x \in [z, y)$, which implies $v_z^* = y$ cannot be optimal, a contradiction.

(b) Consider the case $\gamma \leq y$ and define $\xi = |\gamma - y|$.

(i) If $\gamma - \xi > z$, the argument in (a) can be adapted to show that $v_z^* = y$ cannot be optimal.

(ii) 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. Necessarily, $k < \gamma$ because γ is in the pool, $v_\gamma^* = y$ and $v_s^* = y$ for all minority members $s \in (\gamma, z]$. Therefore, $\mu(x|\{v^*\}) < 1$ and $\iota(x) = 0$. In equilibrium, the individual who expresses x must be located at x (i.e., $k = x$.) Then, however, the individual z in the pool will deviate to $v_z = x$ and obtain the utility $-|z - x| > -|x - y|$, which upsets the equilibrium. Consider now the second possibility, where x is not expressed in equilibrium. In this case, expression of x must be sanctioned: $\mu(x|\{v^*\}) = 1$ (if $\mu(x|\{v^*\}) < 1$, then $v_x^* = x$ and the individual at z will deviate from his equilibrium strategy $v_z^* = y$.) Consider 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 $v_z = k$, contradicting the assumption that in equilibrium, $v_z^* = y$.

(iii) The last case is $\gamma - \xi \leq 0$. By assumption, in any equilibrium $\iota(0) = 0$ (expression of the extreme majority opinion is never sanctioned by the majority.) The individual at z from the pool will deviate to $v_z = 0$ because $-|\gamma - y| \leq -|\gamma - 0|$ implies $-|z - y| < -|z - 0|$, which upsets the pool in equilibrium.

Thus, if equilibrium expression strategies generate a pool, it must be a left-centered pool $\{[z, z'], z\}$ where $y = z$. The next claim establishes 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 first that the belief system must satisfy $\mu(z|\{v^*\}) < 1$ and $\mu(x|\{v^*\}) = 1$ for any $x \in (z, s]$ (if $\mu(z|\{v^*\}) = 1$, then $v_s = s$ yields a higher utility for s, a

contradiction.) The individual located at 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 assumption that $v_s^* = z$ is optimal. Given that any individual located at x expresses an opinion, by monotonicity $v_x^* \leq z$. Since $\mu(z|\{v^*\}) < 1$, the equilibrium strategy and utility of an individual located at $x \in [z, s)$ are respectively $v_x^* = z$ and $-|x - z|$.

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. For such $s, s' \in [0, 1]$, assume, contrary to the claim, that there exists some $y \in (s, s')$ such that the equilibrium strategy of the individual located at y satisfies $v_y^* \neq \emptyset$. 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 $v_x^* \neq \emptyset$ for all $x \in (s, s')$. This is without loss of generality because one can always find infima and suprema of the sets of silent individuals located below and above y.[26]

If (s, s') is a singleton, that is, if this set consists of the point y, we must have $v_y^* = y$ which implies that the strategies of s and s' to remain silent cannot be optimal. Suppose, then, (s, s') is an interval. Using monotonicity and the fact that the open interval (s, s') cannot contain a right-pool or a center-pool, it can be shown that 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 $U_z^* = 0$ and $v_s^* = \emptyset$ 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.

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

Proof. Suppose there exist $s' > s$ such that $v_{s'}^* \in [0, 1]$. Let z be the largest k such that $v_k^* = \emptyset$. By Claim 5, $z < s'$. Now, $v_x^* \in [0, 1]$ for all $x > z$. Consider $x = z + \epsilon$ where $\epsilon > 0$ is sufficiently small. By monotonicity and Claim 3 (no right or centered pool exists) $v_{z+\epsilon}^* = z + \epsilon$. Clearly, for arbitrarily small ϵ, either

[26] That is, define $z = sup\{x|v_x^* = \emptyset, x < y\}$ and $z' = inf\{x|v_x^* = \emptyset, x > y\}$ and let $s = z$, $s' = z'$.

$U_z^* \geq U_{z+\epsilon}^*$ or $U_z^* < U_{z+\epsilon}^*$, implying that either $v_{z+\epsilon}^*$ or v_z^* is not optimal.

Claim 7. *In 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 utility $U_{z_1}^* = U_{z_2}^* = 0$.

Consider an individual between the two pools, $s \in (z_1', z_2)$ arbitrarily close to z_2. By monotonicity, $v_s^* \in (z_1', z_2)$, and by the fact that no left pool exists between the two pools, $v_s^* = s$. Since s is arbitrarily close to z_2, we must have $\mu(s|\{v^*\}) < 1$; the alternative strategy $v_s^* = \emptyset$ cannot be optimal because s can guarantee the payoff $-|z_2 - s|$, which is arbitrarily close to zero, hence larger than $-\alpha$. It follows that $v_x^* = x$ and $\mu(x|\{v^*\}) < 1$. In equilibrium, individuals located at the left neighborhood of x must also be expressing their own opinions. Consider x' in this neighborhood, arbitrarily close to x. Given monotonicity of equilibrium expression strategies, if $v_{x'}^* > x'$, there must exist a right or centered pool, a contradiction. Nor can $v_{x'}^* = \emptyset$ be optimal by the argument above for individual s. Applying this logic successively implies that all $x \in (z_1', z_2)$ must be expressing their own opinions in equilibrium. Then, however, z_1' can increase his utility to arbitrarily close to zero, from $-|z_1' - z_1| < 0$, therefore, $v_{z_1'}^* = z_1$ cannot be optimal, contradicting the assumption that z_1 is the right-extreme member of a left pool.

The results so far establish that equilibrium expression strategies satisfy monotonicity, existence of at most one left pool and that in equilibrium the set of silent individuals, if any, is an interval of the form $[z, 1]$. Given this, the last claim is verified easily.

Claim 8. *There exists a unique expression equilibrium of the form displayed in Proposition 1 or Proposition 2 (up to the sanction F imposed in type-2 equilibria with 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] Bernheim, B. D., (1994) "A Theory of Conformity," *Journal of Political Economy* 102: 841-877.

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

[4] 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.

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

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

[7] 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.

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] Horner, L. R., Conners, J.L. and Daves, R.P. (1998) "Interest in Elections and Public Expression of Opinion," Paper presented at the annual convention of the Midwest Association of Public Opinion Research, Chicago, IL., November.

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

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

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

[20] Lagunoff, R. (2001) "A Theory of Constitutional Standards and Civil Liberty," *Review of Economic Studies* 68, 109-132.

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

[22] 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.

[23] McCroskey, J. C. (1977). "Oral communication apprehension: A summary of recent theory and research." *Human Communication Research* 4: 78-96.

[24] 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.

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] Zaller, J. R., (1992) *The Nature and Origins of Mass Opinion*, Cambridge MA: Cambridge University Press.