Leaders and Social Norms: On the Emergence of Consensus or Conflict

Juan I. Block[a,*], Rohan Dutta[b], David K. Levine[c,d]

[a] *Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom*

[b] *Department of Economics, McGill University, 855 Sherbrooke St. W., Montreal, Quebec, H3A 2T7, Canada*

[c] *Department of Economics, Royal Holloway University of London Egham, Surrey, TW20 0EX, United Kingdom*

[d] *Department of Economics, Washington University in St. Louis, St. Louis MO, 63130-4899, United States of America*

Abstract

We propose a model where competing group leaders influence the social norm adopted in their group constrained by the norm being individually optimal for their members. Individuals are instrumental in enforcing such social norms through peer punishment. We show that there is a unique equilibrium in which there is either a consensus norm or two conflicting norms. A consensus norm is most likely in highly integrated societies, but even in these societies conflicting norms may emerge. Although the majority norm is generally the consensus norm, we characterize the conditions under which the minority norm is adopted as the consensus. In both types of equilibria conformists may not identify with the norm adopted by their group. We show that the intensity of conflict is increasing in the size of the minority group and decreasing in segregation. We also study the welfare and policy implications of our theory.

JEL classification: C72, D71, D74, J15, R23, Z13

Keywords: social norms, leaders, consensus, conflict, peer punishment, collective decision-making

1. Introduction

Social tensions are a common feature of heterogeneous societies in which different groups follow conflicting norms. At the same time, not all heterogeneous societies feature such conflict. In many instances of assimilation (Bazzi et al. (2019), Austen-Smith and Fryer

*Corresponding author

Email addresses: jb2002@cam.ac.uk (Juan I. Block), rohan.dutta@mcgill.ca (Rohan Dutta), david@dklevine.com (David K. Levine)

(2005)), minority groups adopt the majority norm and a consensus emerges. Social norms have been widely studied in the economics literature (Bikhchandani et al. (1992), Bernheim (1994), Lazear (1999), Acemoglu and Jackson (2015), Michaeli and Spiro (2017)), and previous work has focused on, for example, social preferences, interdependence, and information. However, the role of leaders in shaping social norms has received relatively little attention. Our framework builds on the observation that groups invariably have leaders who play a fundamental role in overcoming coordination problems. We study this role of leaders in a tractable model of two groups, each of which must choose a social norm to follow. Our goal is to characterize the conditions under which these environments lead to consensus or conflict and to study the norms adopted by the different groups.

Individuals in our model are either leaders or followers and belong to one of two groups. Each group has a particular norm that is less costly for them to follow than it is for members of the other group. Leaders prefer their own group norm, but members may prefer the other group's norm. This creates a tension between the leaders and the followers. We model the interaction between leaders and followers, and across groups, as a three-stage game.

In the first stage, leaders recommend a social norm for their group members. Our model of group behavior fundamentally differs from the standard approach in that leaders have no coercive power and can only induce their followers to coordinate on a specific social norm provided it is incentive compatible for the followers. Formally, the model is one of a collusion constrained equilibrium, introduced in Dutta et al. (2018). A key feature of the model is that leaders care only about the proportion of the population that adhere to their preferred social norm and so compete over followers.[1]

In a second stage, followers choose a norm. In anticipation of the third stage, they may choose not to follow the norm prescribed by their leader.

In the third and final stage, followers engage in a round of random pairwise social interactions. The chance of a match occurring within a group versus between groups depends on the degree of segregation. Intergroup matches may generate extra benefits and norms are enforced by social sanctions that occur during these matches. Upon being matched, each agent observes the norm chosen by their partner and imposes a punishment on them if it is different from their own.

Our main result characterizes the generically unique collusion constrained equilibrium, which results in either a consensus norm or in two conflicting norms. In a consensus the

[1]The emergence of social norms is often driven by the tension between leaders. For example, the use of contraception promoted by the MCH-FP project in Bangladesh faced strong opposition from local religious leaders (Munshi and Myaux (2006)). Similarly, in many policy implementation decisions, pressure group leaders and/or activists aim to convince people to support their position.

leaders of one group propose their preferred norm, while the leaders of the other group propose the same norm. Everyone adheres to the agreed upon norm. In a conflict, group members adhere to their leaders' preferred norm and punishment occurs when members of different groups interact.

We use this framework to derive new testable hypotheses and explore the effect of social factors (for example, segregation and diversity) on the role of leaders driving consensus or conflict. We find that consensus emerges at low levels of segregation, but, nonetheless, conflicting norms will emerge even in fully integrated societies if the minority is large enough. In particular, conflict occurs whenever the minority share exceeds a threshold, which depends on the level of segregation. Similarly, conflict arises whenever the degree of segregation exceeds a threshold, which depends on the relative group sizes and also on the net costs to a member from following the opposing group's norm instead of their own. The intensity of conflict, as measured by the expected peer punishment, decreases with greater segregation and increases with the minority group size.

We also study the types of norms that are followed in equilibrium. Our theory, for instance, supports the common view that the minority group members may be induced by social pressure to adhere to a majority norm that they do not like, which we refer to as "coerced." However, our theory also predicts that if both groups prefer the minority norm, then that norm can become the consensus. Here we should emphasize that our assumption is not that a group prefers their own norm to the other, but rather that their own norm is more desirable to them than it is for the other group. In this sense, it is the opposite case, where both groups prefer the minority norm, but never-the-less the equilibrium is the majority norm, that is surprising: this occurs when the majority is substantially larger than the minority.

Our theoretical framework also sheds light on the impact of leaders on welfare. The presence of the leaders can have either positive and negative welfare implications for group members. A marginal increase in segregation reduces welfare in all scenarios but one. In the exception, a society originally in conflict and with a costly enough punishment, benefits from greater segregation through fewer instances of intergroup matches. We conclude by discussing how introducing simple dynamics into our model generates the phenomenon of tipping as discussed in Schelling (1971).

1.1. Related Literature

Our work contributes to the literature that seeks to understand the emergence of common or conflicting norms when individuals enforce social norms through peer pressure (Munshi and Myaux (2006), Michaeli and Spiro (2015; 2017), Henry and Louis-Sidois (2020), among

others). As in that literature, individuals trade off costs of social pressure, due to miscoordination, against personal benefits, due to having intrinsic preferences over norms. However, unlike these studies, we explicitly model how groups choose social norms. Our model provides novel insights into the mechanisms of norm selection within groups, highlighting the role of leadership and the impact of equilibrium norms adoption on group welfare. Our contribution is to explore the interaction between groups that are organized by leaders who are able to coordinate their followers' actions through tools (such as peer pressure and ostracism), whose efficacy depends on the choices made by other groups. In this we follow the classic works of Olson (1965) and Ostrom (1990), and more recently, Levine and Modica (2016; 2020). In contrast, Kets and Sandroni (2021) identify an alternative driver of group coordination, modeling culture as shared cognition. The key insight of their model is that cultural diversity reinforces strategic uncertainty.

This paper complements the theoretical literature on leadership that emphasizes the informational role of leaders (De Mesquita (2010), Acemoglu and Jackson (2015), Shadmehr and Bernhardt (2019), Chen and Suen (2021), Morris and Shadmehr (2023)). Leaders in these models have better information about key parameters, and, in equilibrium, followers partly infer this information from their leader's actions. Unlike these papers, leaders in our model have organizational role and limited ability to coordinate their group members. Leaders in our model do not possess superior information.

There are several related papers that study heterogeneous societies where individuals benefit the most when they interact with those who adhere to the same norm, and often find multiplicity of equilibria (Bikhchandani et al. (1992), Bernheim (1994), Lazear (1999), Mengel (2008), Kuran and Sandholm (2008), Advani and Reich (2015), Bazzi et al. (2019)). In contrast to these papers, we find that conformism is not necessarily beneficial for the individuals. Like Kets and Sandroni (2021), our model provides equilibrium selection when individuals interact in environments where coordination is crucial; however, we do not assume group members follow an introspection process and face impulses.

2. Model

2.1. Environment

A society consists of two groups $J \in \{A, B\}$. There is a continuum of individuals of unit mass, with a fraction $0 < \phi_A < 1$ who are members of group A, and with the remaining fraction $\phi_B = 1 - \phi_A$ being members of group B. In addition to group members, each group has leaders of infinitesimal mass.

There are two social norms $j \in \{a, b\}$. A social norm is a code of conduct, such as tax compliance, customs, and traditions, and so on. Social norms are group specific in that

norms a and b correspond to group A and B, respectively.

Group members have intrinsic preferences over norms. For any member of group J adhering to the social norm k has an individual cost of c_{Jk}. These costs could take negative values, thereby representing benefits. This is the personal payoff associated with a social norm choice. We assume each member of a group likes their own social norm better than members of the other group do.

Assumption 1. $c_{Aa} \leq c_{Ba}$ *and* $c_{Bb} \leq c_{Ab}$.

This assumption permits only one of two scenarios. Either one norm is commonly preferred by all group members or members of each group prefer their own norm. The first scenario, if a is the commonly preferred norm, requires $c_{Aa} \leq c_{Ab}$ and $c_{Ba} \leq c_{Bb}$. The second scenario requires instead that $c_{Aa} \leq c_{Ab}$ and $c_{Bb} \leq c_{Ba}$. Importantly, the assumption rules out the possibility that each group prefers the other group's norm. The first scenario may apply when one social norm is harmful. For example, early female marriage is associated with lower schooling and domestic violence for young women as well as more rapid spread of disease across communities (Field and Ambrus (2008)). Alternatively, one social norm may benefit both groups; for example, smoking in public spaces being generally unaccepted can have a positive effect on smokers' and nonsmokers' health.

The leaders of each group specify simultaneously and independently the social norm that should be followed by each member of their group. Each individual takes as given the norm chosen by everyone else in society (including fellow members) and adheres to the norm specified by their own leader provided it is in their interest to do so. The leaders of each group prefer their own social norm to that of the other group. That is, even if members of both groups agree on one norm, the leaders do not. This can be thought as, for example, the political leaders' ideology in partisan conflicts, or managers who endorse different work practices (Akerlof and Kranton (2000)). In particular, leaders do not necessarily have the same preferences as their members. Leaders compete to impose their preferred norm: their objective function is the fraction of the population that adheres to their preferred social norm.[2] In the case of political leaders, this fraction of the population can be interpreted as power in the form of supporters and voters that in turn would result in a higher likelihood of being elected. In the case of organizations, the proportion of workers adopting a particular

[2]These preferences are standard in that the leaders want the aggregate behavior to comport with their preference and suffer a loss resulting from departures of individuals' choices from their preferred outcome. One such example is the leader's loss function assumed in Dewan and Myatt (2008, 2012). Assuming that leaders obtain some benefit when their group members follow their proposed norm yet adhere to the other group's norm has no qualitative effect on our results as long as this benefit is strictly less than when followers adhere to the group's norm.

work practice may deliver more profits to the managerial unit. If their followers are unwilling to adhere to the norm they propose then we assume that the leader suffers a ruinous utility loss. This could be thought as leaders' accountability in that they would be replaced if they proposed a norm that is not optimal for their members.

After the social norms are determined by the leaders, group members engage in a round of social interaction. Specifically, individuals are matched randomly in pairs: with probability $1-\sigma$ the entire population is matched randomly, and with probability σ each group is matched randomly with own group members only. We refer to σ as the *degree of segregation*.

Upon being matched, each member observes whether the matched partner adhered to the same social norm or not. Social norms are assumed to rely on *peer enforcement* by which individuals must penalize deviations from accepted behavior. We assume that there is a fixed punishment $P > 0$ that is imposed by a group member on a partner who fails to comply. This may be in the form of informal social sanctions such as peer pressure and social ostracism, or other kinds of physical or material sanctions.

We study situations where peer punishment is sufficient to enforce social norm compliance. To do so, we assume the cost of being punished is greater than the cost of switching social norms.

Assumption 2. *For any group J and norms j, k with $j \neq k$, $P > |c_{Jk} - c_{Jj}|$.*

In addition to inflicting social pressure, outgroup interactions are known to deliver additional benefits, for example, by offering a different perspective or skill (Lazear (1999), Hong et al. (1998), and Alesina et al. (2000)). We assume that a member who meets a member of the other group obtains a payoff U (which need not be positive).[3] Putting these pieces together, the (expected) payoff for a member of group $J \in \{A, B\}$ who chooses norm $j \in \{a, b\}$ is given by

$$\pi_{Jj} = (1-\sigma)(1-\phi_J)U - c_{Jj} - \mu_{Jj}P,$$

where μ_{Jj} is the probability of meeting a partner adhering to a different norm. Notice that even though P is a constant, the expected punishment for an individual, $\mu_{Jj}P$, can vary with μ_{Jj}. The latter depends both on the degree of segregation, an exogenous variable, and the proportion of the individual's group that adheres to the group norm, an equilibrium variable.

[3]We assume that the effects of diversity is driven by the identity of the groups and not by the adopted social norms per se (see, for example, Alesina et al. (2016)).

The following parameter captures a notion of norm identification,

$$d_J \equiv \frac{c_{Jk} - c_{Jj}}{P},$$

where j is group J's norm and $k \neq j$.

If $d_J > 0$ then group J members have positive norm identification and identify with their group's norm. If $d_J < 0$ then they have negative norm identification and identify with the other group's norm. The magnitude of d_J captures the cost (or benefit) relative to being punished, for group J members who follow the opposing social norm $k \neq j$.

Assumptions 1 and 2 are reflected in the following key properties of d_J:

Lemma 1. *$d_A + d_B \geq 0$ and $-1 < d_A, d_B < 1$.*

We end this section with the following genericity assumption:

Assumption 3. *For each group J, $\phi_J \neq \dfrac{1 - 2\sigma - d_J}{2(1-\sigma)}$.*

2.2. Equilibrium

We have assumed leaders have a limited ability to specify the norm in the sense that group members will only adopt the social norm proposed by their leaders if it is incentive compatible. Hence, leaders will only choose such norms. Specifically, a norm is incentive compatible for a group if no group member can be made better off by following a different norm given that everyone else in the group follows the norm. Crucially, whether a norm is incentive compatible for a group depends on the actions of the other group. The notion of equilibrium that captures this idea is collusion constrained equilibrium (CCE).[4] In the current context, the set of CCE would be identical to the prediction from the following simpler equilibrium notion, which for the sake of brevity we continue to refer to as CCE.

Definition 1. A *collusion constrained equilibrium in the social norm game (CCE)* is a choice of a social norm by the leaders of each group such that, given the choice of the leaders of the other group, it is incentive compatible for members to adhere to the norm and no other incentive compatible norm is preferred by either leader.

If in equilibrium leaders of both groups choose the same social norm we refer to *consensus*, and if in equilibrium leaders of each group choose their preferred social norm we refer to *conflict*.

[4]See Dutta et al. (2018) for a formal justification for using this solution concept when studying interaction between groups. CCE applies broadly to any non-cooperative game in which the players are partitioned into collusive groups, and is defined in an appropriately subtle way to avoid non-existence problems.

3. Consensus and Conflict

In this section, we characterize the conditions that lead to a consensus or conflict, and which social norm is adopted when there is consensus.

Our main result shows that generically there is a unique collusion constrained equilibrium, and sharply characterizes parameter configurations for which it features consensus, or conflict.

Proposition 1. *If $\phi_J > \frac{1+d_K}{2(1-\sigma)}$ with $K \neq J$, then there is a unique collusion constrained equilibrium with consensus on group J's norm. Otherwise, there is a unique collusion constrained equilibrium with conflict.*

Our formal proof is in the Appendix A; here we discuss the idea. Note that if both norms are incentive compatible for a group, the group leaders would strictly prefer to propose their preferred social norm. Then we need to characterize the conditions under which it is incentive compatible for the group members to adhere to their leaders' preferred norm given the other group's behavior. To this end, it is optimal for group J members to adhere to (their own) social norm while the other group members follow their own social norm if the population share ϕ_J is above the following threshold

$$\underline{\phi}_J(\sigma, d_J) \equiv \frac{1 - 2\sigma - d_J}{2(1-\sigma)}.$$

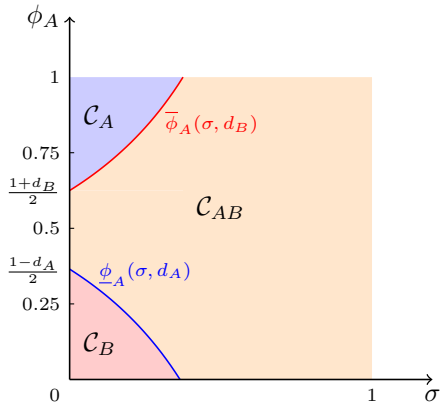
It is incentive compatible for members in group $K \neq J$ to adhere to the group J social norm when the population share ϕ_J is above the threshold

$$\overline{\phi}_J(\sigma, d_K) \equiv \frac{1 + d_K}{2(1-\sigma)}.$$

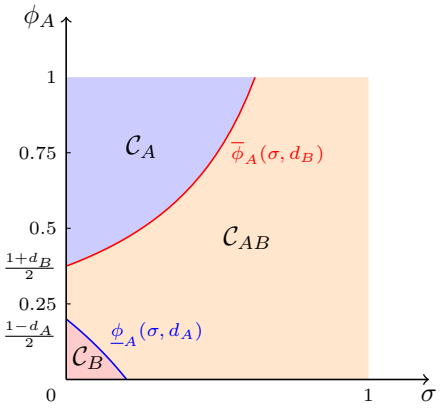
Observe that $1 > \overline{\phi}_J \geq \underline{\phi}_J > 0$, by Lemma 1, and that $\overline{\phi}_J = 1 - \underline{\phi}_K$ with $K \neq J$. These thresholds are sufficient to describe the collusion constrained equilibrium as described in the Proposition.

Figure 1 illustrates the conflict and consensus regions in the unique equilibrium described in Proposition 1. The two Panels in Figure 1 differ only on switching costs. Given d_A, d_B, the *conflict region* $\mathcal{C}_{AB}$ in equilibrium based on (σ, ϕ_A) is defined by

$$\mathcal{C}_{AB}(d_A, d_B) \equiv \left\{(\sigma, \phi_A) \mid \underline{\phi}_A(\sigma, d_A) < \phi_A < \overline{\phi}_A(\sigma, d_B)\right\},$$



(a) Group members identify with their norm, $d_J > 0$ for all J.



(b) Both group members identify with norm a, $d_B < 0 < d_A$.

Figure 1: Consensus and conflict as functions of σ and ϕ_A.

and the *consensus on group J's social norm region* $\mathcal{C}_J$ in equilibrium is given by[5]

$$\mathcal{C}_J(d_K) \equiv \left\{(\sigma, \phi_A) \mid \phi_J > \overline{\phi}_J(\sigma, d_K)\right\}.$$

The Role of Leaders. Removing leaders from the social norm game, effectively makes it a static game wherein individuals opt for the norm they prefer given the choices of others. In this case, the relevant equilibrium notion is Nash equilibrium. To evaluate the role of leaders, it is therefore instructive to discuss the relationship between CCE and standard Nash equilibrium in the social norm game. By Assumption 2, if the social norm game did not involve choice by the leaders, for any configuration of parameters, there would always be Nash equilibria in which everyone adheres to the same social norm (i.e., consensus). The only relevant deviations in such Nash equilibria, by definition, are those of individuals, holding everyone else's choice fixed, both ingroup and outgroup.

Leaders, in this paper, can coordinate the choice of their group members on any social norm so long as it is incentive compatible given the behavior of those outside the group. The requirement of ingroup incentive compatibility is the same as the Nash equilibrium requirement of no profitable deviation for any individual. The novelty here is that the coordinating ability of the leader makes group deviations relevant too. Holding fixed the outgroup behavior, the ingroup could effectively deviate to a different incentive compatible norm following their leader. This additional requirement of CCE, as Proposition 1 shows, is

[5]Note that, by Assumption 2, in equilibrium the consensus regions are non-empty. This assumption implies that, eventually, the leaders of at most one group would fail to make their preferred norm incentive compatible through ingroup peer punishment.

sufficient to rule out all Nash equilibria but one. This unique equilibrium may entail conflict or consensus depending on parameter values.

The uniqueness result also relies on the specific preferences of the leaders. If the leaders were indifferent between the two norms, then equilibrium multiplicity would persist.

Equilibrium Social Norms. Each of the two types of equilibria, consensus and conflict, features different behavioral implications that depends on the parameters. A consensus, for instance, may be on the majority or the minority norm. Furthermore, in equilibrium, a group may follow the costlier norm, through what may be interpreted as *coercion*.

If the members of each group exhibit positive norm identification (see Figure 1a), then a consensus equilibrium must feature the majority norm. In this case, the minority group is assimilated and adheres to the norm even though it has higher cost to them. This occurs, for example, when both a minority group of immigrants and the citizens of the host country strongly identify with their respective group social norms. Unlike Kuran and Sandholm (2008), consensus here does not generate social tensions. By contrast, in a conflict equilibrium group members identify with their ingroup conforming norm. For instance, Bisin et al. (2008) find that Muslims in the UK do not integrate even when they reside in mixed neighborhoods (see the part of region $\mathcal{C}_{AB}$ in Figure 1a consistent with low segregation).

Suppose the majority group prefers the minority norm to their own, which we refer to as "negative norm identification," and that segregation is sufficiently low. If the majority group is not too large (see the part of region $\mathcal{C}_A$ in Figure 1b that lies below $\phi_A = 1/2$), the minority norm may be adopted in a consensus equilibrium. None of the groups is coerced. This kind of assimilation is not possible in the model proposed by Advani and Reich (2015). On the other hand, if the majority is sufficiently large (see region $\mathcal{C}_B$ in Figure 1b), there is consensus over a *biased norm*, that is, a norm that is privately rejected by all individuals. Our result offers a novel explanation for the drinking behavior observed in college settings, where leadership roles are particularly influential, and students may privately feel discomfort with excessive alcohol consumption (Prentice and Miller (1993)). Additionally, we identify the leaders of majority groups as a mechanism through which biased norms persist, distinct from the mechanism proposed by Michaeli and Spiro (2017).

If one group shows negative norm identification (by assumption at most one group can do so), when there is conflict only one group is coerced and its members adhere to the group norm due to ingroup peer punishment (see region $\mathcal{C}_{AB}$ in Figure 1b). This prediction points out a different interpretation of results found in Bazzi et al. (2019) that segregated communities are more likely to identify strongly with their own ethnic group. According to our model, in segregated communities, one would observe ethnic group members adhering to a common norm induced by their leaders; however, they may not necessarily identify

with such norm. This is also consistent with the peer effect observed when Black community leaders (the minority group) impose costs on their members who try to "act White" (Austen-Smith and Fryer (2005)), whereby Black students decide not to conform to the majority norm of grade achievements and academic effort.

3.1. Segregation, Minority Group Size, Norm Identification, and Punishment

Proposition 1 has clear implications about the different economic and social parameters that map to consensus or conflict. We discuss these in turn.

Segregation

An increase in segregation lowers the punishment cost of adhering to the leaders' preferred norm while increasing the punishment cost of violating it, irrespective of what the other group is doing and the group members' intrinsic preferences. So, if the equilibrium is conflict at a given set of parameter values, then it will remain so at a higher level of segregation. Further, for any pair of population shares, there will be conflict at a sufficiently high level of segregation. This follows from the observations that the thresholds $\overline{\phi}_J$ and $\underline{\phi}_J$ are increasing and decreasing in σ, respectively, and that if $\sigma > (1-\min_{J\in\{A,B\}}\{d_J\})/2$, then there is conflict regardless of group sizes. The segregation threshold at which the equilibrium switches from consensus to conflict may be higher when negative norm identification is present.

Similar to the theoretical results of Kuran and Sandholm (2008), Proposition 1 suggests that, regardless of population shares, more segregated societies are more likely to have groups adhering to conflicting norms. This is consistent with the findings in Corvalan and Vargas (2015) on the positive effect of ethnic and language segregation on the incidence of civil conflicts at any intensity level. This prediction is shared with other models (e.g., Kuran and Sandholm (2008)), however, our theory provides a novel explanation why segregation is often correlated with conflict. Segregation enhances ingroup peer punishment which, in turn, allows leaders to coordinate group members on their preferred social norm. Empirically, this result points out that segregation can be the main driving force behind conflict. For instance, Corvalan and Vargas (2015) find that several conflict-related variables (for example, indices based on group sizes) were not significant when segregation was included in models estimating the relation between conflict and segregation.

As Figure 1 shows, the minimum level of segregation required for conflict is higher when one group dominates in size. Importantly, our theory emphasizes that the relation between segregation and conflicting norms can be weak. For example, fellow members generally identify with their religious group norm (they display positive norm identification), and consequently for intermediate group sizes segregation may have no effect on the onset of

conflicting norms (see region $\mathcal{C}_{AB}$ in Figure 1a). In the same vein, Corvalan and Vargas (2015) find no significant effect of religious segregation on conflict.

We complete our discussion of segregation with a thought experiment. If leaders could influence the degree of segregation, what would they do? Since in a consensus equilibrium leaders of one group fail to see their preferred norm being followed by any individual, such leaders would have an incentive to increase segregation among groups; however, the leaders of the mainstream norm would advocate the opposite. An example of this dynamic is when minority leaders advocate segregation in response to assimilation pressures (Austen-Smith and Fryer (2005)). If we allowed leaders to influence segregation (at least marginally), then the losing leader would always aim to segregate their group members while the winning leader would foster integration. Therefore, the overall effect is unclear.

Minority Group Size

The minority group size (which fully describes the population shares in our model), is an important parameter in determining the equilibrium norm(s). As previously discussed, because segregation eventually allows conflicting norms in equilibrium, we analyze the size of the minority group when segregation is sufficiently low (i.e., $\sigma < (1 - \max\{d_A, d_B\})/2$). So long as the population distribution is consistent with consensus, with one exception, a sufficiently large increase in the population share of the minority induces the minority group to switch and adopt their own group's norm. This leads to conflicting norms in equilibrium. This is because increasing the relative population share of the minority group makes its peer enforcement stronger. The exception arises when members of both groups prefer the minority group leaders' norm and this is also the only candidate for consensus due to the small size of the majority group (see region $\mathcal{C}_A$ in Figure 1b with $\overline{\phi}_A \leq \phi_A < 1/2$). Here increasing the relative population share of the minority group makes the minority group leaders' norm even more attractive for the majority group, leaving the consensus norm unchanged.

Our model also provides an alternative explanation to the empirical findings in Echenique and Fryer (2007). They find that when black students are relatively scarce in schools, they tend to integrate quite well (see region $\mathcal{C}_B$ in Figure 1a). However, as their group size increases, they do not integrate, and the critical threshold is approximately 25 percent of the school population. Our result can help better understand why students adopt conflicting norms as the minority group surpasses a threshold.

Norm Identification

Greater identification for one's own group norm is reflected in higher values of d_J. As can be seen from contrasting Figures 1a and 1b, higher d_J increases the range of parameters σ and ϕ_A where conflict occurs. If it was an equilibrium to follow your own group's norm,

regardless of what the other group does, then increasing your preference for that norm would only reinforce the equilibrium. Indeed, such an equilibrium can then emerge with lesser segregation than previously required. This observation can be stated formally as

$$d_J \leq d'_J \Rightarrow \mathcal{C}_{AB}(d_J, d_K) \subseteq \mathcal{C}_{AB}(d'_J, d_K).$$

This observation suggests that when individuals strongly identify with their group norm, societies are more likely to exhibit conflicting norms, as in Advani and Reich (2015). In the UK, Muslims exhibit a strong religious identity and integrate more slowly than other minority groups (Bisin et al. (2008)). The summer of 2024 saw a surge in extreme right-wing idealogical activity, fueling faith-based hate crimes and social tensions in the UK. These events illustrate how strong group norm identification often leads to conflict.

Punishment Severity

We now study the effect of the punishment P on the predictions of the model. When individuals identify with their own group norm, the harsher the punishment is, the more difficult is for the minority group leaders to impose their preferred norm. Figures 1a and 1b show how the conflict region $\mathcal{C}_{AB}(d_A, d_B)$ shrinks when P increases since $\overline{\phi}_J$ shifts downward and $\underline{\phi}_J$ move outwards.

When all individuals identify with group J's norm, increasing P improves the ability of the leaders to coordinate on their preferred norm depending on their group size; thus, the overall impact of this parameter change is less clear. Regardless of the relative size of their group, a harsher punishment is instrumental for the $K \neq J$ leaders (whose group members do not identify with their preferred norm) by allowing more severe ingroup peer punishment. Looking at Figure 1b, one can see that an increase in P and thus a decrease in d_J, shrinks the conflict region by increasing $\underline{\phi}_J(\sigma, d_J)$. On the other hand, it makes it more difficult for the J leaders to impose their preferred norm when their group is the minority. In Figure 1b this would be represented by a shift of the curve $\overline{\phi}_J(\sigma, d_K)$ to the right.

Unlike Spiro (2020), who finds that increasing social sanctioning unequivocally reduces extreme behavior in equilibrium, our model suggests that this outcome depends critically on the extent to which group members identify with their group's norm. This identification introduces ambiguity in the impact of social sanctions on consensus and conflict.

At the limit, when punishment becomes arbitrarily harsh ($P \to \infty$), the role of norm identification is erased ($d_A, d_B \to 0$). The degree of segregation and relative population shares are all that matter. At this limit Figure 1 becomes symmetric with the graphs of the functions $\overline{\phi}_A$ and $\underline{\phi}_A$ meeting at $\phi_A = 1/2$ in the absence of segregation ($\sigma = 0$). In such a fully integrated society, there is always consensus on the majority group norm.

4. Intensity of Conflict

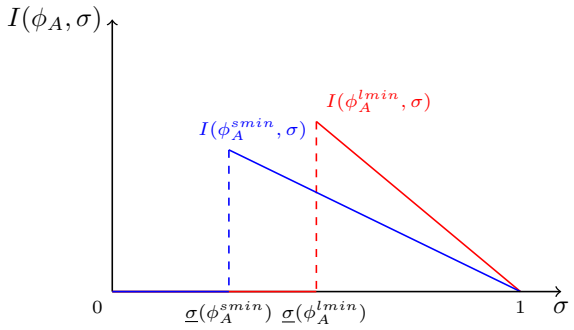
In our model different conflict equilibria typically yield different amounts of conflict because punishment occurs only when agents adhering to different social norms interact. With almost complete segregation, for example, there is conflict only in the hypothetical sense that if anyone actually met they would punish the partner. Given a conflict equilibrium, the relevant measure of the level of conflict is therefore the expected cost of punishment per capita

$$I(\phi_A, \sigma) = (1 - \sigma)\phi_A\phi_B P,$$

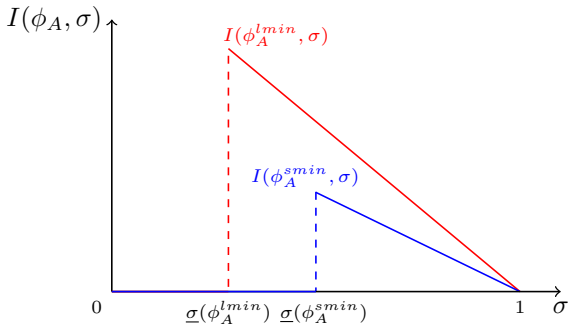
which we label the *intensity of conflict*. Conditional on conflict, the intensity of conflict is decreasing in the degree of segregation σ and increasing in the minority size.

Starting from a consensus equilibrium, where there is no conflict, so intensity is zero, increasing segregation eventually triggers the switch to conflict and intensity jumps up. At this point the intensity of conflict is at its maximum. Further segregation now dampens the intensity. This is because despite hostile intent (opposing norms) the two groups meet less and less often.

This result is consistent with the evidence of Field et al. (2008) that incidents of violence were more likely to occur in integrated neighborhoods in the 2002 riots in India.



(a) All group members identify with norm b, $d_A < 0 < d_B$.



(b) Group members identify with their norm, $d_J > 0$ for all J.

Figure 2: Intensity of conflict with respect to σ.

The change in intensity of conflict with respect to segregation is shown in Figure 2 for different levels of population shares and norm identification. The symbol ϕ_A^{lmin} corresponds to a society with a large minority, while ϕ_A^{smin} represents one with a small minority. Given d_A, d_B and ϕ_A, the lowest level of segregation consistent with equilibrium conflicting social norms is

$$\underline{\sigma}(\phi_A, d_A, d_B) = \max\left\{0, \frac{1 - 2\phi_A - d_A}{2(1 - \phi_A)}, \frac{2\phi_A - 1 - d_B}{2\phi_A}\right\}.$$

Panel (a) corresponds to the case where both groups prefer norm b so an increase in the size of the minority increases the lowest level of segregation consistent with conflict. Panel (b) captures the more standard case in which a larger minority group requires less segregation to generate conflict. Notice though that in both cases, if the level of segregation is consistent with conflict at both sizes of the minority group, the intensity of conflict is higher with a larger minority group. In other words, more diverse societies would yield more costly conflicts if ever initiated.

From a policy perspective, this finding suggests that, starting from a highly segregated society, interventions that foster integration could result in more severe conflicts. This is similar to the results obtained from intergenerational culture transmission models (Buechel et al. (2014), Spiro (2020)). For example, Spiro (2020) finds that the more influential the groups are on each other, the slower the process of reaching a common norm. In our model, the influence of the other group can be thought as how likely members from different groups would interact, which is captured by the degree of segregation. Another policy implication of this result is that larger resources should be devoted to integration programs if the minority group is large.

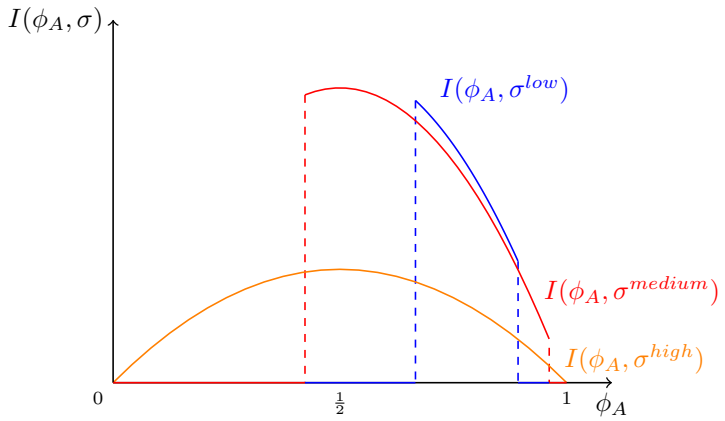


Figure 3: Intensity of conflict with respect to ϕ_A.

Figure 3 offers a different perspective by mapping the intensity of conflict as a function of ϕ_A for different degrees of segregation, $\sigma^{low} < \sigma^{medium} < \sigma^{high}$. Our finding that intensity is increasing in the minority size is consistent with the results of most of the literature on the economics of conflict (for example, Esteban and Ray (2011; 2012), Esteban et al. (2012)). However, it highlights the key difference: while low levels of segregation allow for conflict only over a more limited range of population distributions, when conflict does occur, it is more intense. Furthermore, the intensity of conflict does not necessarily attain its maximum when the groups have equal size.

5. Welfare Analysis

In this section, we explore welfare in the collusion constrained equilibrium. Some relevant welfare comparisons are unambiguous and admit a Pareto ranking. For others it is useful to specify a cardinal measure of welfare. In particular, we work with the average expected payoff under conflict and consensus j, respectively, defined by

$$W_{\mathcal{C}_{AB}}(\phi_A, \sigma) = 2(1-\sigma)\phi_A\phi_B(U-P) - \phi_A c_{Aa} - \phi_B c_{Bb}, \tag{1}$$

$$W_{\mathcal{C}_J}(\phi_A, \sigma) = 2(1-\sigma)\phi_A\phi_B U - \phi_A c_{Aj} - \phi_B c_{Bj}. \tag{2}$$

5.1. Assessing Group Leaders

To evaluate the welfare implications of leaders, we need to model the counterfactual decentralized outcome. This is captured by Nash equilibrium behavior in the social norm game without leaders. In particular, the members of each group simultaneously choose a norm and it is optimal for each member given the choice of the others. As mentioned earlier, there is a multiplicity of such Nash equilibria. Both group members following the same norm is a Nash equilibrium, irrespective of the parameters. Further, the unique CCE, which captures the role of leaders, is itself one such Nash equilibrium. We do not take any stand on which Nash equilibrium would be selected in the decentralized setting. Instead we compare the unique CCE to the remaining set of decentralized Nash equilibria.

Since the CCE is itself a Nash equilibrium it is straightforward that the equilibrium outcome in the presence of leaders can also emerge as a decentralized equilibrium. What is unclear is how the CCE compares with the other Nash equilibria.

We first explore the situation in which all individuals intrinsically prefer a particular social norm. The next result identifies when leaders induce the best possible aggregate welfare.

Proposition 2. *If both group members prefer social norm j, the welfare of group members of each group is weakly larger with leaders than without, only if there is a consensus CCE on norm j.*

When all individuals intrinsically prefer a single norm, consensus on that norm (decentralized or via leaders) always yields higher welfare for members of each group. In the example of Black community leaders rejecting acting white (conflict CCE), Austen-Smith and Fryer (2005) find that a salient welfare consequence to their members is that Black students obtain worse economic outcomes in the job market.

Proposition 3. *Suppose both group members identify with their group norm. In a consensus CCE on group J norm the welfare of group $K \neq J$ members is weakly smaller with leaders*

than without, and the opposite holds for group J members. In a conflict CCE the welfare of members of both groups can be smaller or larger with leaders than without.

5.2. Segregation

This section explores how the degree of segregation σ affects welfare in a CCE. As discussed in Section 3.1, a marginal increase in segregation can lead to three possible scenarios. A consensus equilibrium remains a consensus equilibrium, a conflict equilibrium remains as such and finally a consensus equilibrium switches to conflict. In the next proposition we summarize the impact on welfare in these three cases.

Proposition 4. *Suppose there is a marginal increase in segregation σ.*

(i) At a consensus CCE, if the type of equilibrium is unchanged, then welfare strictly decreases if $U > 0$, strictly increases if $U < 0$, and is constant otherwise.

(ii) At a conflict CCE welfare decreases if and only if $U \geq P$.

(iii) If the CCE switches from consensus to conflict, then welfare decreases.

Parts (i) and (ii) follow immediately from equations (1) and (2). If intergroup meetings generate a net surplus, then clearly greater segregation reduces welfare. For part (iii), notice that for the group whose norm was the consensus, say J, a move to conflict brings the penalty P from being matched with the other group, $K \neq J$. The latter faces the same penalty but now may face a lower cost from following their own norm. Nevertheless, at the point where the equilibrium switches, it must be that following their own norm is weakly incentive compatible for K members. But then, their welfare in a conflict equilibrium is their welfare in the consensus equilibrium decreased by the outgroup punishment.

Separatism and Integration. There has been a surge of secessionism in developing countries (Morelli and Rohner (2015)) as well as in Western democracies (Gehring and Schneider (2018)). Separatists often base their arguments on cultural/nation identity and political autonomy, and the idea that the group would benefit from separation. Unionists, on the other hand, argue that those becoming independent would be worse off by losing access to some markets or facing public good provision issues. There is also a heated debate about whether religious groups are associated with segregated lifestyle and radical views, or they enhance the diversity of societies. There is a vast literature on the costs and benefits of secession (see, for example, Alesina and Spolaore (2005)); here, we focus on the individuals' costs and benefits of adhering to social norms.

In the context of our theory, this normative question corresponds to asking whether a conflict equilibrium can ever generate greater welfare than a consensus. Proposition 4 shows

that to answer this it is sufficient to compare consensus without segregation $W_{\mathcal{C}_J}(\phi_A, 0)$ to conflict with total segregation $W_{\mathcal{C}_{AB}}(\phi_A, 1)$.

Proposition 5. *Suppose ϕ_A is consistent with consensus on group J norm for low enough segregation. Then, $W_{\mathcal{C}_{AB}}(\phi_A, 1) \geq W_{\mathcal{C}_J}(\phi_A, 0)$ if and only if $d_K P \geq 2\phi_J U$ for $K \neq J$.*

Intuitively, for conflict with total segregation to generate higher welfare the consensus norm must be costly enough for the group with the other norm to outweigh the payoff U from a complete lack of segregation. This result says that secession may lead to a welfare improvement as long as one group complies with a consensus norm that it does not identify with and strongly dislikes. Furthermore, if both groups prefer the consensus norm, then intergroup interactions must necessarily incur a cost to individuals for secession to be welfare-improving.

Some immigrant religious groups do not seem to integrate in their host country, even after spending there several years (Bisin et al. (2008)). To tackle this issue, many Western countries have implemented policies that are designed to restrict religious expression and foster integration, such as the 2004 French headscarf ban and the 2023 French Abaya ban. Our result predicts that these policies may either result in the assimilation of religious groups (the intended outcome) or in more intense conflict. In both cases the effect on welfare is ambiguous. Abdelgadir and Fouka (2020) find that, on average, the educational outcomes and economic integration of Muslim women was negatively affected by the law. Our result also implies that if such religious groups were to identify with the mainstream social norm, but their leaders do not, then (full) integration could be welfare improving. Incidentally, Abdelgadir and Fouka (2020) show that the negative effect of the ban was mitigated for women who readily identify with French values.

6. Discussion

Dynamics. We have examined the relationship between segregation and the choice of norms in a static model. In this model it is possible to see also the implications of certain dynamics. Suppose, in particular, that conflicting norms lead to greater segregation. In this case Figure 1 confirms that once in conflict, such a society would enter a cycle of increasing segregation and persistent conflict, each reinforcing the other. It is not necessary, though, that conflict would then lead quickly to a totally segregated society. Recall that the intensity of conflict decreases with segregation. If segregation is increasing in the intensity of conflict, then our model would predict a slowing down of segregation over time. We would expect to see societies caught in a conflict-segregation cycle but far from complete segregation.

Schelling (1971) discusses the phenomenon of tipping wherein a minority group enters a neighborhood in sufficient numbers causing the majority residents to begin evacuating. The key feature is a critical threshold for the minority share, a tipping point, below which not much changes and above which the original majority residents eventually all leave. Card et al. (2008) find evidence of tipping behavior in a number of US cities, with tipping points ranging from 5% to 20% minority share. Our model coupled with the simple dynamic in the paragraph above generates tipping behavior. Assuming A to be the majority group, a society with initial segregation σ would have a tipping point of $\underline{\phi}_B(\sigma, d_B) = 1 - \overline{\phi}_A(\sigma, d_B)$. In our theory it is the minority group's choice of norm, rather than its mere presence, that determines the dynamics of segregation. Interestingly, the tipping point depends on the preferences of the minority and (perhaps more surprisingly) not on that of the majority. The rationale is that the distaste for conflict is what persuades the majority to move. The minority share threshold for conflict, that is above which the minority stop adopting the majority norm and instead hold their own, is wholly determined by the preferences of the minority.

Parameters U and P. We use the parameter P to model the punishment at the heart of the peer enforcement mechanism. The exogenously set fixed value specification of P delivers a tractable model and transparent analyses. Future work could explore a variety of other natural specifications, such as one that varies depending on the identity of agents, or the size of their group. These specifications seem more compelling if we interpret the parameter more generally as the cost of miscoordination.

We use the parameter U to capture any additional benefit or cost that flows from interactions across groups. As can be seen above, the parameter plays no role whatsoever in the findings in sections 3 and 4. Unsurprisingly it matters in the welfare analysis of section 5. We do not take a stand on its sign, and report the results for all cases.

Acknowledgments

We thank Matt Elliott, Andrea Mattozzi, and Salvatore Modica. Financial support from the EUI Research Council, The Leverhulme Trust and the Janeway Institute is gratefully acknowledged.

References

Abdelgadir, A. and Fouka, V. (2020). Political secularism and muslim integration in the west: Assessing the effects of the french headscarf ban. *American Political Science Review*, 114(3):707–723.

Acemoglu, D. and Jackson, M. O. (2015). History, expectations, and leadership in the evolution of social norms. *Review of Economic Studies*, 82(2):423–456.

Advani, A. and Reich, B. (2015). Melting pot or salad bowl: the formation of heterogeneous communities. Technical report, IFS Working Papers.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753.

Alesina, A., Harnoss, J., and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, 21(2):101–138.

Alesina, A. and Spolaore, E. (2005). *The Size of Nations*. MIT Press.

Alesina, A., Spolaore, E., and Wacziarg, R. (2000). Economic integration and political disintegration. *American Economic Review*, 90(5):1276–1296.

Austen-Smith, D. and Fryer, R. G. (2005). An economic analysis of "acting white". *Quarterly Journal of Economics*, 120(2):551–583.

Bazzi, S., Gaduh, A., Rothenberg, A. D., and Wong, M. (2019). Unity in diversity? how intergroup contact can foster nation building. *American Economic Review*, 109(11):3978–4025.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5):841–877.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Bisin, A., Patacchini, E., Verdier, T., and Zenou, Y. (2008). Are muslim immigrants different in terms of cultural integration? *Journal of the European Economic Association*, 6(2-3):445–456.

Buechel, B., Hellmann, T., and Pichler, M. M. (2014). The dynamics of continuous cultural traits in social networks. *Journal of Economic Theory*, 154:274–309.

Card, D., Mas, A., and Rothstein, J. (2008). Tipping and the dynamics of segregation. *Quarterly Journal of Economics*, 123(1):177–218.

Chen, H. and Suen, W. (2021). Radicalism in mass movements: Asymmetric information and endogenous leadership. *American Political Science Review*, 115(1):286–306.

Corvalan, A. and Vargas, M. (2015). Segregation and conflict: An empirical analysis. *Journal of Development Economics*, 116:212–222.

De Mesquita, E. B. (2010). Regime change and revolutionary entrepreneurs. *American Political Science Review*, 104(3):446–466.

Dewan, T. and Myatt, D. P. (2008). The qualities of leadership: Direction, communication, and obfuscation. *American Political Science Review*, 102(3):351–368.

Dewan, T. and Myatt, D. P. (2012). On the rhetorical strategies of leaders: Speaking clearly, standing back, and stepping down. *Journal of Theoretical Politics*, 24(4):431–460.

Dutta, R., Levine, D. K., and Modica, S. (2018). Collusion constrained equilibrium. *Theoretical Economics*, 13(1):307–340.

Echenique, F. and Fryer, R. G. (2007). A measure of segregation based on social interactions. *Quarterly Journal of Economics*, 122(2):441–485.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and conflict: An empirical study. *American Economic Review*, 102(4):1310–42.

Esteban, J. and Ray, D. (2011). Linking conflict to inequality and polarization. *American Economic Review*, 101(4):1345–74.

Esteban, J.-M. and Ray, D. (1994). On the measurement of polarization. *Econometrica*, pages 819–851.

Field, E. and Ambrus, A. (2008). Early marriage, age of menarche, and female schooling attainment in bangladesh. *Journal of Political Economy*, 116(5):881–930.

Field, E., Levinson, M., Pande, R., and Visaria, S. (2008). Segregation, rent control, and riots: The economics of religious conflict in an indian city. *American Economic Review*, 98(2):505–10.

Gehring, K. and Schneider, S. A. (2018). Towards the greater good? eu commissioners' nationality and budget allocation in the european union. *American Economic Journal: Economic Policy*, 10(1):214–39.

Henry, E. and Louis-Sidois, C. (2020). Voting and contributing when the group is watching. *American Economic Journal: Microeconomics*, 12(3):246–76.

Hong, L., Page, S. E., et al. (1998). Diversity and optimality. Santa Fe Institute.

Kets, W. and Sandroni, A. (2021). A theory of strategic uncertainty and cultural diversity. *Review of Economic Studies*, 88(1):287–333.

Kuran, T. and Sandholm, W. H. (2008). Cultural integration and its discontents. *Review of Economic Studies*, 75(1):201–228.

Lazear, E. P. (1999). Culture and language. *Journal of Political Economy*, 107(S6):S95–S126.

Levine, D. K. and Mattozzi, A. (2020). Voter turnout with peer punishment. *American Economic Review*, 110(10):3298–3314.

Levine, D. K. and Modica, S. (2016). Peer discipline and incentives within groups. *Journal of Economic Behavior & Organization*, 123:19–30.

Mengel, F. (2008). Matching structure and the cultural transmission of social norms. *Journal of Economic Behavior & Organization*, 67(3-4):608–623.

Michaeli, M. and Spiro, D. (2015). Norm conformity across societies. *Journal of Public Economics*, 132:51–65.

Michaeli, M. and Spiro, D. (2017). From peer pressure to biased norms. *American Economic Journal: Microeconomics*, 9(1):152–216.

Morelli, M. and Rohner, D. (2015). Resource concentration and civil wars. *Journal of Development Economics*, 117:32–47.

Morris, S. and Shadmehr, M. (2023). Inspiring regime change. *Journal of the European Economic Association*, 21(6):2635–2681.

Munshi, K. and Myaux, J. (2006). Social norms and the fertility transition. *Journal of Development Economics*, 80(1):1–38.

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard Economic Studies.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Prentice, D. A. and Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64(2):243.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2):143–186.

Shadmehr, M. and Bernhardt, D. (2019). Vanguards in revolution. *Games and Economic Behavior*, 115:146–166.

Spiro, D. (2020). Multigenerational transmission of culture. *Journal of Economic Theory*, 188:105037.

Appendix A. Proofs

Proof of Lemma 1. Write $P(d_A + d_B) = c_{Ab} - c_{Aa} + c_{Ba} - c_{Bb} = (c_{Ab} - c_{Bb}) + (c_{Ba} - c_{Aa})$. By Assumption 1 and $P > 0$ it follows $d_A + d_B \geq 0$. By Assumption 2, for any group J and social norms $j \neq k$, $-P < c_{Jk} - c_{Jj} < P$; $-1 < d_J < 1$ follows by definition. □

Proof of Proposition 1. Suppose all norms are incentive compatible for both group members. The payoffs to the group leaders from the two choices of social norm are given by

		B leaders	
		a	b
A leaders	a	$1, 0$	ϕ_A, ϕ_B
	b	ϕ_B, ϕ_A	$0, 1$

Observe that for the leaders their own social norm strictly dominates the other group's social norm.

We next study incentive compatibility for group members. The expected payoff of a group J member adhering to norm j is given by $\pi_{Jj} = (1-\sigma)(1-\phi_J)U - c_{Jj} - \mu_{Jj}P$, where μ_{Jj} is the probability of meeting a partner adhering to a different norm. By Assumption 2, if both groups follow a social norm, it is optimal for everyone to do so. If

$$\pi_{Jj} > \pi_{Jk} \iff (1-\sigma)(1-\phi_J)P < d_J P + (\sigma + (1-\sigma)\phi_J)P,$$

it is strictly incentive compatible for members of group J to adhere to their group social norm $j \neq k$ even if the members of the other group $K \neq J$ do not and strictly not incentive compatible when the inequality is reversed strictly. This is without loss of generality by Assumption 3. Rewrite this as

$$\phi_J > \frac{1 - 2\sigma - d_J}{2(1-\sigma)} \equiv \underline{\phi}_J(\sigma, d_J). \tag{A.1}$$

If this is the case then the leaders of group J will choose their group's social norm j as this is their most preferred norm.

If inequality (A.1) holds for J leaders and the opposite for $K \neq J$ leaders, namely, the following condition is satisfied

$$\phi_K < \frac{1 - 2\sigma - d_K}{2(1-\sigma)}, \tag{A.2}$$

then J leaders will choose their own social norm and K leaders will have no choice but to conform, resulting in the consensus on group J social norm. The latter, inequality (A.2), may be rewritten using $\phi_K = 1 - \phi_J$ as

$$\phi_J > \frac{1 + d_K}{2(1-\sigma)} \equiv \overline{\phi}_J(\sigma, d_K). \tag{A.3}$$

By Lemma 1, since $d_K \geq -d_J$, we have

$$\frac{1 + d_K}{2(1-\sigma)} \geq \frac{1 - 2\sigma - d_J}{2(1-\sigma)},$$

so that if inequality (A.3) holds so does inequality (A.1). Hence consensus is the unique

equilibrium when inequality (A.3) holds for one of the two groups.

By Assumption 3, there are two other possibilities. If both group leaders' dominant strategies are incentive compatible then there is a unique equilibrium where they follow these strategies resulting in conflict. Alternatively, none of the group leaders choosing their own social norm in the face of their opponents choosing theirs is incentive compatible for their members. The theorem follows from ruling out this latter possibility. We show that at least leaders of one group are able to implement their preferred norm in the face of the other leaders doing the same.

Suppose that it is not feasible for J leaders to implement their own social norm in the face of $K \neq J$ leaders implementing their group social norm. From reversing inequality (A.1) and by Assumption 3 this requires that

$$\phi_J < \frac{1 - 2\sigma - d_J}{2(1-\sigma)}.$$

Using $\phi_J = 1 - \phi_K$ and $d_J \geq -d_K$ this can be written as

$$\phi_K > \frac{1 - 2\sigma - d_K}{2(1-\sigma)},$$

which implies that it is feasible for group K's leaders to implement their own social norm even when leaders of group J implements their own social norm. □

Proof of Proposition 2. Assume, without loss of generality, $d_B < 0 < d_A$. In a conflict CCE the welfare of each group $J \in \{A, B\}$ is given by

$$\omega^J_{\mathcal{C}_{AB}} = (1-\sigma)\phi_A\phi_B(U - P) - \phi_J c_{Jj}.$$

A Nash equilibrium with consensus on norm a, the welfare of group A is given by

$$\omega^A_a = (1-\sigma)\phi_A\phi_B U - \phi_A c_{Aa}$$

and for group B is given by

$$\omega^B_a = (1-\sigma)\phi_A\phi_B U - \phi_B c_{Ba}.$$

It follows that $\omega^J_a > \omega^J_{\mathcal{C}_{AB}}$ for each group $J \in \{A, B\}$. □

Proof of Proposition 3. Suppose $d_J > 0$ for both $J \in \{A, B\}$. Consider a consensus CCE on norm j. The welfare of group J in this consensus CCE is

$$\omega_{C_J}^J = (1-\sigma)\phi_A\phi_B U - \phi_J c_{Jj}.$$

In a Nash equilibrium with consensus on norm $k \neq j$, the welfare of group J is given by

$$\omega_k^J = (1-\sigma)\phi_A\phi_B U - \phi_J c_{Jk}.$$

Hence

$$\omega_{C_J}^J - \omega_k^J > 0 \Longleftrightarrow \phi_J d_J P > 0.$$

Comparing the welfare of group $K \neq J$ under both scenarios we obtain that

$$\omega_{C_J}^K - \omega_k^K < 0 \Longleftrightarrow -d_K P\phi_K < 0.$$

In a conflict CCE the welfare of each group $J \in \{A, B\}$ is given by

$$\omega_{\mathcal{C}_{AB}}^J = (1-\sigma)\phi_J\phi_K(U-P) - \phi_J c_{Jj}.$$

By Proposition 2, it suffices to consider a Nash equilibrium with consensus on the other group's norm k. The welfare of group J is given by

$$\omega_k^J = (1-\sigma)\phi_A\phi_B U - \phi_J c_{Jk},$$

Then

$$\begin{aligned}
\omega_{\mathcal{C}_{AB}}^J - \omega_k^J &= (1-\sigma)\phi_J\phi_K(U-P) - \phi_J c_{Jj} - (1-\sigma)\phi_J\phi_K U + \phi_J c_{Jk} \\
&= \phi_J(c_{Jk} - c_{Jj}) - (1-\sigma)\phi_J\phi_K P \\
&= \phi_J d_J P - (1-\sigma)\phi_J\phi_K P
\end{aligned}$$

Therefore, the welfare of group J is larger in a conflict CCE than in a Nash equilibrium with consensus on norm k if

$$\begin{aligned}
\omega_{\mathcal{C}_{AB}}^J - \omega_k^J > 0 &\Longleftrightarrow d_J - (1-\sigma)\phi_K > 0 \\
&\Longleftrightarrow \phi_K < \frac{d_J}{1-\sigma}.
\end{aligned}$$

And, the welfare of group $K \neq J$ is larger in a conflict CCE than in a Nash equilibrium with

consensus on norm k if

$$\begin{aligned}\omega^K_{\mathcal{C}_{AB}} - \omega^K_j > 0 \Longleftrightarrow & d_K - (1-\sigma)\phi_J > 0 \\ \Longleftrightarrow & \phi_K > \frac{1-\sigma-d_K}{1-\sigma}.\end{aligned}$$

We know that for a conflict equilibrium to exist we need that

$$\frac{1-2\sigma-d_J}{2(1-\sigma)} < \phi_K < \frac{d_J}{2(1-\sigma)}.$$

□

Proof of Proposition 4(iii). Without loss of generality, assume the consensus equilibrium was a. Consider the welfare difference

$$W_{\mathcal{C}_A}(\phi_A, \sigma) - W_{\mathcal{C}_{AB}}(\phi_A, \sigma) = -\phi_B(c_{Ba} - c_{Bb}) + 2(1-\sigma)\phi_A\phi_B P.$$

For this to be positive requires

$$P > \frac{c_{Ba} - c_{Bb}}{2(1-\sigma)\phi_A}.$$

Since we are evaluating this inequality at the point where the equilibrium switches from consensus to conflict, we must set $\phi_A = \overline{\phi}_A(\sigma, d_B) = (1+d_B)/(2(1-\sigma))$. Substituting this above gives

$$P > \frac{c_{Ba} - c_{Bb}}{1+d_B}.$$

Recall that $d_B = (c_{Ba} - c_{Bb})/P$. So we have

$$1 > \frac{c_{Ba} - c_{Bb}}{P + c_{Ba} - c_{Bb}},$$

which is always satisfied since $P > 0$. □

Proof of Proposition 5. We prove the statement for $j = a$. A symmetric argument applies to the other case.

$$\begin{aligned}& W_{\mathcal{C}_{AB}}(\phi_A, 1) \geq W_{\mathcal{C}_A}(\phi_A, 0) \\ \Longleftrightarrow & -\phi_A c_{Aa} - \phi_B c_{Bb} \geq 2\phi_A\phi_B U - \phi_A c_{Aa} - \phi_B c_{Ba} \\ \Longleftrightarrow & c_{Ba} - c_{Bb} \geq 2\phi_A U \\ \Longleftrightarrow & d_B \geq (2\phi_A U)/P.\end{aligned}$$

□