Leaders and Social Norms: On the Emergence of Consensus or Conflict[1]

Juan I. Block[2], Rohan Dutta[3], David K. Levine[4]

Abstract

We propose a model where competing leaders influence the social norm adopted in their group provided it is optimal for their members. Individuals are instrumental in enforcing such social norm through peer punishment. We show that there is a unique equilibrium in which there is either a consensus norm or conflicting norms. A consensus norm is most likely in highly integrated societies, but nevertheless conflicting norms may emerge in fully integrated societies. Although the majority norm is generally the consensus norm, we characterize the conditions under which the minority norm is adopted. 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 minority group size and decreasing in segregation. We also study welfare and policy implications of our theory.

JEL Classification Numbers: C72, D71, D74, J15, R23, Z13

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

[1]We thank Andrea Mattozzi, and Salvatore Modica. Financial support from the EUI Research Council and the Cambridge-INET Institute is gratefully acknowledged.

[2]Faculty of Economics, University of Cambridge; email: jb2002@cam.ac.uk

[3]Department of Economics, McGill University; email: rohan.dutta@mcgill.ca

[4]Department of Economics and RSCAS at the EUI and WUSTL; email: david@dklevine.com

1. Introduction

Social tensions are a common feature of heterogeneous societies in which the different groups follow conflicting norms. At the same time, not all heterogeneous societies feature such conflict. As in numerous instances of assimilation, minority groups often adopt the majority norm and consensus emerges. Groups invariably have leaders and are influenced by these prominent agents. Leaders coordinate group members' expectations and shape their behaviors by, for example, their actions or words. We propose a tractable leadership-based model to study why consensus emerges in some societies while others must contend with conflicting norms, and to study what kinds of norms are conformed by different groups.

Individuals in our model are either leaders or followers and belong to one of two groups. Each group has a corresponding norm, one that is easier (less costly) to follow for the group members compared to those in the other group and that is preferred by the group leaders. Importantly, members may not identify with their group norm, that is, they may prefer the other group's norm. We model the interaction between leaders and followers, and across groups, as a three-stage game.

In the first stage, leaders specify a social norm for their group members. Our model of group behavior fundamentally differs from the standard approach in that leaders can 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 are competing. This assumption aims to capture that the emergence of social norms is often driven by the tension between status quo and opposition 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)). During the Covid-19 pandemic, there has been significant heterogeneity among health experts and political leaders regarding the role of social distancing in preventing the spread of the virus.

In a second stage, followers choose a norm. They may not undiscerningly follow the norm prescribed by their leaders because they know that in a third stage they will be matched with an agent and would feel social pressure when behaving differently from each other. In the third and last stage, followers engage in a round of social pairwise random interactions. The chance of a match occurring within a group versus between groups depends on the degree of segregation. While intergroup matches may generate extra benefits, norms are enforced by social sanctions. 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. Punishments allow leaders to impose their preferred norm if their group is completely segregated; however,

intergroup interactions may hinder their ability to do so.

Our main result characterizes the generically unique equilibrium, which results in either a consensus or conflict. In a consensus the leaders of only one group impose their preferred norm, which the leaders of the other group, abandoning their own norm, also propose. Everyone adheres to the norm. In a conflict, group members adhere to their leaders' preferred norm and punishment occurs when members of different groups interact.

We find that consensus emerges at low levels of segregation, but conflicting norms may coexist in fully integrated societies. When groups are more likely to interact, the size of the groups and whether individuals identify with their group norm shape members incentive to adhere to the norms. Conflict obtains 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 groups norm instead of their own. Groups that are similar in size lowers this segregation threshold as long as members identify with their group norm. The intensity of conflict, measured by the expected peer punishment, decreases with greater segregation and increases with the minority group size.

In addition to characterizing when a common norm is adhered by all groups, we study what types of norms are followed and who uphold the consensus norm in equilibrium. Our theory, for instance, supports the common view that the minority group is often alienated (that is, they adhere to a norm that they dislike) if the majority norm is the mainstream norm and groups identify with their group norm. But our theory makes the less common prediction that the minority norm can be the prevailing norm in society; however, in this case the majority group must identify with such norm in order to adopt it. We also find that when groups adhere to conflicting norms either the majority or the minority group, or neither, can be alienated.

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 mechanically 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 paper focuses on the role of leaders in influencing the adoption of social norms in a setting where groups may be in conflict. Acemoglu and Jackson (2015) argues that leaders are able to determine which social norms emerge in equilibrium by affecting how overlapping generations interpret past information. We model leaders as hierarchical, yet not coercive,

agents that can coordinate their fellow group members' behavior so long as it is strategically optimal for their followers.

More broadly, our work contributes to the growing 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). Our contribution is to explore the interaction between groups that are able to coordinate their actions through tools (peer pressure, 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) and Levine and Mattozzi (2020). Some papers study heterogeneous societies where individuals benefit the most when they interact with those who adhere to the same norm, and find multiplicity of equilibria (Lazear (1999), Mengel (2008), Advani and Reich (2015), Bazzi et al. (2019)). In contrast to these papers, we find a unique equilibrium with sharp comparative statics results that can be fundamental in forming public policy. The paper is also related to recent experimental studies that have demonstrated that group behavior has positive effects, increasing prosocial behaviors towards ingroup members (Goette et al. (2006), Charness et al. (2007), Chen and Li (2009)), as well as negative effects, in the form of costly conflict between groups, generating hostility, and antisocial actions, toward outsiders (Goette et al. (2012)). Our framework builds upon a growing empirical literature that has shown that leaders are able to influence individuals both positively and negatively (Kosfeld and Rustagi (2015)). See, for example, Beekman et al. (2014), d'Adda et al. (2017), Ajzenman (2021) for evidence on leaders promoting dishonest norms among citizens, and see, for instance, Andreoni (2006), Jack and Recalde (2015), Güth et al. (2007) on leaders inducing them to increase their contributions to public goods. We contribute to this literature by providing a set of testable predictions in settings where leaders from different groups have opposing goals.

2. Model

2.1. Environment

Consider a society consisting 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 dress codes, customs, and traditions, and so on, shared by group members and enforced through

social sanctions.[5] Social norms are group specific in that norm j corresponds to group J. 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. We assume each member of a group likes their own social norm better than members of the other group do.

Assumption 1. *$c_{Jj} \leq c_{Kj}$ for $K \neq J$.*

However, notice that members of one group may prefer the other group's social norm, that is we allow $c_{Jj} > c_{Jk}$ for $k \neq j$. We allow this because one social norm may be 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, security of tenure to tenants and larger share of output that is paid as rent on farm productivity to landlords can have a positive effect on agricultural productivity (Banerjee et al. (2002)).

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 group J prefer their own social norm j to that of the other group $k \neq j$. Leaders are competing: their objective function is the fraction of the population that adheres to their preferred social norm. If their followers are unwilling to adhere to the norm they propose then the leader suffers a ruinous utility loss.

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. In addition to inflicting social pressure, outgroup interactions may deliver additional benefits, for example, by offering a different perspective or skill (see Hong et al. (1998) and Alesina et al. (2000)). We assume

[5] We note that the punishment of non-adherents is a characteristic of social norms: it appears, for example, even in the written constitution of a prison gang (see, for example, Skarbek (2014)). For theoretical consideration, see Levine and Modica (2016).

that a member who meets a member of the other group obtains a benefit $U \geq 0$.

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 may choose only such social norms. A norm is incentive compatible for a group if no group-member can better off by following a different norm while 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).[6] 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*.

We require that the punishments be large enough to induce compliance with social norms. To this end, we assume that the cost of being punished is greater than the cost of switching social norms.

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

To avoid special cases due to group members being ex ante indifferent, we make the following genericity assumption:

Assumption 3. *For each group J and $k \neq j$, $\phi_J \neq \dfrac{1-2\sigma}{2(1-\sigma)} - \dfrac{(c_{Jk} - c_{Jj})/P}{2(1-\sigma)}$.*

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.

[6] 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.

Define $d_J \equiv (c_{Jk} - c_{Jj})/P$ as the net cost relative to being punished for group J members who adhere to the opposing social norm $k \neq j$. Our assumptions about c_{Jk} 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 can then interpret the parameter d_J as *norm identification* and distinguish between three cases. Group J members have positive norm identification if $d_J > 0$; they identify with their group norm. Whereas they have negative norm identification if $d_J < 0$; they identify with the norm of the other group. Finally, they have neutral norm identification if $d_J = 0$. The closer $|d_J|$ is to 1, the stronger (negative or positive) norm identification is. The lemma also says that at least members of one group have positive norm identification.

Our main result shows that generically there is a unique collusion constrained equilibrium.

Proposition 1. *If $\phi_J > \dfrac{1 + d_{-J}}{2(1-\sigma)}$, there is a unique collusion constrained equilibrium with consensus j. 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 j while the other group members follow social norm $k \neq j$ 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 $-J$ to adhere to the norm j when the population share ϕ_J is above the threshold

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

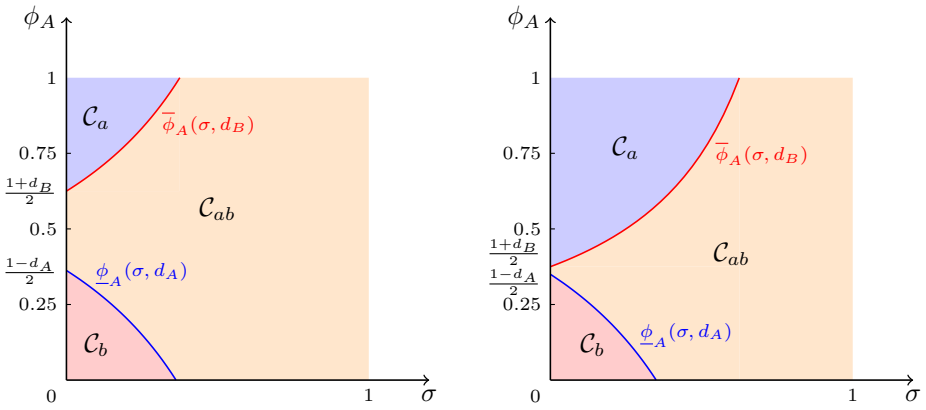
Observe that $\overline{\phi}_J \geq \underline{\phi}_J$ by Lemma 1 and that $\overline{\phi}_J = 1 - \underline{\phi}_{-J}$. These thresholds are sufficient to describe the collusion constrained equilibrium as described in the Proposition.

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

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

and the consensus j region $\mathcal{C}_j$ in equilibrium is given by

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



(a) $d_J > 0$ for all J. (b) $d_B < 0 < d_A$.

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

3.1. Segregation, Minority Group Size and Norm Identification

We next analyze how the emergence of consensus and conflict depend on the economic environment.

Segregation. Greater segregation implies less interaction between groups, making peer enforcement within groups stronger. As a result, the required incentive to adhere to the leaders' preferred norm may be achieved with a smaller group size, as the following observation states.

Corollary 1. *The thresholds $\overline{\phi}_J$ and $\underline{\phi}_J$ are increasing and decreasing in σ, respectively.*

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. So, for sufficiently high segregation two conflicting norms would have to be an equilibrium, as the following corollary shows.

Corollary 2. *If $\sigma > (1 - \min\{d_J\})/2$, there is conflict regardless of group sizes.*

This result suggests that regardless of population shares, more segregated societies are more likely to have groups adhering to conflicting norms.[7] As Figure 1 makes clear, the relationship between segregation and conflict is monotone, conflict at a given level of segregation

[7]This is consistent with the findings in Corvalan and Vargas (2015) on the effect of ethnic and language segregation on the incidence of civil conflicts at any intensity level.

implies conflict at all higher levels. Indeed we can characterize the level of segregation at which the equilibrium (if ever) at consensus switches to conflict.

Proposition 2. *The lowest level of segregation consistent with equilibrium conflict as a function of population shares and switching costs, $\underline{\sigma}(\phi_A, d_A, d_B)$, satisfies*

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

The graph of $\underline{\sigma}(\cdot, d_A, d_B)$ as a function of ϕ_A can be seen in Figure 1 as the curve separating the consensus and conflict regions. As the figure shows this minimum level of segregation required for conflict is higher when one group dominates in size.

Minority Group Size. We next turn to the role played by the size of the minority group in the emergence of consensus and conflict. The relationship between the size of the minority group and conflict is not straightforward in that a given population share of the minority group may or may not be consistent with conflict depending on switching costs, size of the majority group and degree of segregation. To describe the relationship indirectly we characterize the effect of the minority on the lowest level of segregation consistent with equilibrium conflict. A decreasing (increasing) effect means that a larger minority encourages conflict (consensus).

Proposition 3.

(i) If $d_J > 0$ for both $J \in \{A, B\}$, then $\underline{\sigma}(\phi_A, d_A, d_B)$ is strictly decreasing in the minority group size for $\phi_J > \overline{\phi}_J$.

(ii) If $d_K < 0$ and $J \neq K$, then $\underline{\sigma}(\phi_A, d_A, d_B)$ is decreasing in the minority group size for $\phi_J \geq 1/2$ or $\phi_J \leq \underline{\phi}_J$ and increasing in the minority group size for $\overline{\phi}_J \leq \phi_J < 1/2$.

(iii) If $\underline{\phi}_J < \phi_J < \overline{\phi}_J$ then $\underline{\sigma}(\phi_A, d_A, d_B) = 0$ (and therefore independent of the minority group size).

In words, so long as the population distribution is consistent with consensus at some level of segregation then, with one exception, an increase in the population share of the minority lowers the minimum level of segregation that supports conflict. Increasing the relative population share of the minority group makes its peer enforcement stronger and therefore less reliant on segregation. The exception arises when members of both groups prefer the minority group leaders' norm, which is also the only candidate for consensus due to the small size of the majority group (see 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, thereby requiring even greater segregation to make the majority group stick to their own norm.

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, now such an equilibrium could emerge with lesser segregation than was earlier required. The following result states this observation formally.

Proposition 4. *The conflict region $\mathcal{C}_{ab}(d_A, d_B)$ is monotonic in d_J, i.e. if $d_J \leq d'_J$ then $\mathcal{C}_{ab}(d_J, d_{-J}) \subseteq \mathcal{C}_{ab}(d'_J, d_{-J})$.*

This result implies that when individuals have a strong identification with their group norm, societies are more likely to exhibit conflicting norms.

3.2. Equilibrium Social Norms

We characterize two types of equilibria, yet these may imply very different behavioral predictions in our setting. These may differ in the fundamental feature of whether individuals identify with the conforming group norm. Furthermore, consensus equilibria vary in terms of who effectively uphold the common norm.

If the members of each group exhibit positive norm identification (that is, $d_J > 0$), then a consensus equilibrium must feature the majority norm. In this case, the minority group is alienated, that is, they adhere to the norm that they dislike, and the consensus norm is supported by the majority group. By contrast, in a conflict equilibrium group fellows identify with their ingroup conforming norm.

Suppose the majority group displays negative norm identification and that segregation is low (that is, $\sigma < (1 - \min\{d_J\})/2\}$). If the majority group is not too large, its norm may not be adopted in a consensus equilibrium. None of the groups is alienated and, interestingly, the consensus norm is upheld by the minority group. On the other hand, if the majority is sufficiently large, there is consensus over a biased norm, that is, a norm that is privately rejected by all individuals. Most importantly, the majority group who dislike the consensus norm are the ones upholding it, influenced by their leaders.

On the other hand, if either the majority or the minority group shows negative norm identification, when there is conflict only one group is alienated and its members adhere to the norm due to ingroup peer punishment. This is consistent with the peer effect observed when Black communities (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, and consequently they obtain worse economic outcomes in the job market.

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

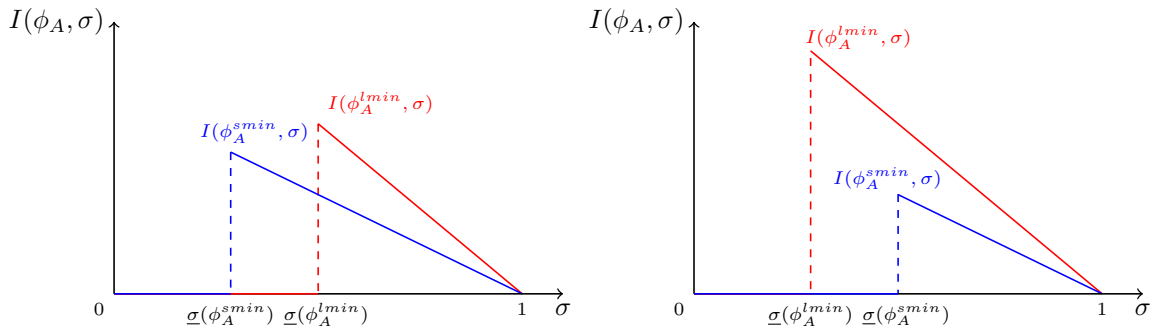
$$I(\phi_A, \sigma) = (1 - \sigma)\phi_A\phi_B P,$$

which we label the *intensity of conflict*. The next proposition simply lists which parameters influence this intensity, and how and follows directly from the equation above.

Proposition 5. *Conditional on conflict, the intensity of conflict is decreasing in the degree of segregation σ, increasing in $\phi_A\phi_B$ and independent of switching costs.*

Starting from a consensus equilibrium, increasing segregation eventually triggers the switch to conflict. At this point the intensity of conflict is at its maximum. Further segregation only dampens the intensity since despite clear hostile intent (opposing norms) the two groups meet less and less often. The result is also 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.

The change in intensity of conflict with respect to segregation is shown in Figure 2 for different levels of population shares. ϕ_A^{lmin} corresponds to a society with a large minority and ϕ_A^{smin} to one with a small minority. Panel (a) corresponds to the exceptional case in proposition 3 part (ii) where 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.



(a) $d_A < 0 < d_B$. (b) $0 < d_A < d_B$.

Figure 2: Intensity of conflict with respect to σ.

Figure 3 offers a different perspective by mapping the intensity of conflict as a function of ϕ_A for different degrees of segregation, $\sigma^{low} < \sigma^{medium} < \sigma^{high}$. It highlights the key observation that while low levels of segregation allow for conflict for a more limited range of population distributions, when it does it generates more intense conflicts.

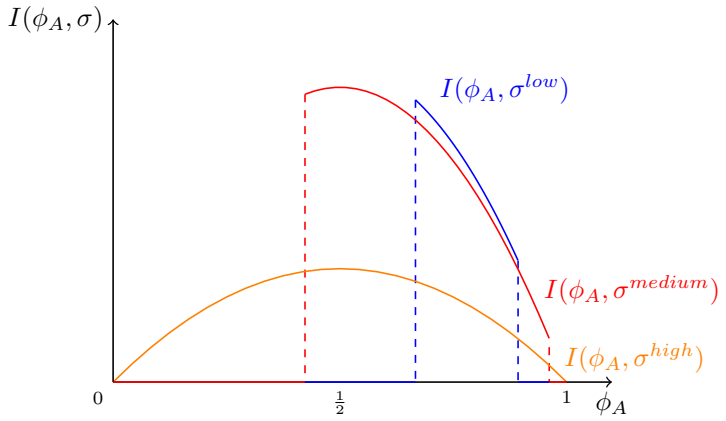


Figure 3: Intensity of conflict with respect to ϕ_A.

5. Segregation and Welfare

This section explores how the degree of segregation affects welfare in equilibrium. The average expected payoff under conflict and consensus j, respectively, are

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

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

As discussed in Subsection 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 6. *Suppose there is a marginal increase in segregation.*

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

(ii) At a conflict equilibrium if the type of equilibrium is unchanged then welfare decreases if and only if $U \geq P$.

(iii) If the equilibrium 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, $-J$. 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 $-J$ members. But then, their welfare in a conflict equilibrium is their welfare in the consensus equilibrium decreased by the outgroup punishment.

5.1. 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.

In the context of our theory, this normative question corresponds to asking whether a conflict equilibrium can ever generate greater welfare compared to consensus. Proposition 6 shows that to answer this it is sufficient to compare consensus without segregation $W_j(\phi_A, 0)$ to conflict with total segregation $W_{ab}(\phi_A, 1)$.

Proposition 7. *Suppose* ϕ_A *is consistent with consensus* j *for low enough segregation. Then,* $W_{ab}(\phi_A, 1) \geq W_j(\phi_A, 0)$ *if and only if* $d_{-J} \geq 2\phi_A U/P$.

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 benefit 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 they do no identify with and strongly dislikes.

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. Our result predicts that these policies may result in more intense conflict or alienation of religious groups, thereby decreasing welfare. Abdelgadir and Fouka (2020) find that 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 would 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

We have examined the relationship between segregation and the choice of norms in a static model. Nevertheless it is easy to see 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 to a totally segregated society in a hurry. 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 sufficiently 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 above which the minority stop adopting the majority norm and instead hold their own, resulting in conflict, is wholly determined by the preferences of the minority.

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.

Ajzenman, N. (2021). The power of example: Corruption spurs corruption. *American Economic Journal: Applied Economics*, 13(2):230–57.

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

Andreoni, J. (2006). Leadership giving in charitable fund-raising. *Journal of Public Economic Theory*, 8(1):1–22.

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

Banerjee, A. V., Gertler, P. J., and Ghatak, M. (2002). Empowerment and efficiency: Tenancy reform in west bengal. *Journal of Political Economy*, 110(2):239–280.

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.

Beekman, G., Bulte, E., and Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural liberia. *Journal of Public Economics*, 115:37–47.

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.

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

Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, 97(4):1340–1352.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

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

d'Adda, G., Darai, D., Pavanini, N., and Weber, R. A. (2017). Do leaders affect ethical conduct? *Journal of the European Economic Association*, 15(6):1177–1213.

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

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.

Goette, L., Huffman, D., and Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, 96(2):212–216.

Goette, L., Huffman, D., Meier, S., and Sutter, M. (2012). Competition between organizational groups: Its impact on altruistic and antisocial motivations. *Management science*, 58(5):948–960.

Güth, W., Levati, M. V., Sutter, M., and Van Der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, 91(5-6):1023–1042.

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.

Jack, B. K. and Recalde, M. P. (2015). Leadership and the voluntary provision of public goods: Field evidence from bolivia. *Journal of Public Economics*, 122:80–93.

Kosfeld, M. and Rustagi, D. (2015). Leader punishment and cooperation in groups: Ex-

perimental field evidence from commons management in ethiopia. *American Economic Review*, 105(2):747–83.

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.

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.

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

Skarbek, D. (2014). *The social order of the underworld: How prison gangs govern the American penal system*. Oxford University Press.

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, $-P < c_{Jk} - c_{Jj} < P$; $-1 < d_J < 1$ 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

$$
\begin{array}{cccc}
 & & B \text{ leaders} & \\
 & & a & b \\
A \text{ leaders} & a & 1,0 & \phi_A, \phi_B \\
 & b & \phi_B, \phi_A & 0,1
\end{array}
$$

Observe that for the leaders of J their own social norm j strictly dominates $k \neq j$.

We next study incentive compatibility for group members. The expected payoff of a group J member adhering to norm j is given by $\pi_{Jj} = -c_{Jj} - \mu_{Jj}P + (1-\sigma)\phi_{-J}U$, 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

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

it is incentive compatible for members of J to adhere to j even if $-J$ members do not and strictly not incentive compatible when the inequality is reversed. 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 J will choose j as this is their most preferred norm.

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

$$
\phi_{-J} < \frac{1-2\sigma-d_{-J}}{2(1-\sigma)}, \tag{A.2}
$$

then J leaders will choose j and $-J$ leaders will have no choice but to conform, resulting in the consensus j. The latter, inequality (A.2), may be rewritten using $\phi_{-J} = 1 - \phi_J$ as

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

By Lemma 1, since $d_{-J} \geq -d_J$, we have

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

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 $-J$ leaders implementing $-j$. From reversing inequality (A.1) and by Assumption 3 this requires that

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

Using $\phi_J = 1-\phi_{-J}$ and $d_J \geq -d_{-J}$ this can be written as

$$\phi_{-J} > \frac{1-2\sigma-d_{-J}}{2(1-\sigma)},$$

which implies that it is feasible for $-J$ to implement $-j$ even when J implements j. □

Proof of Proposition 6(iii). Without loss of generality, assume the consensus equilibrium was a. Consider the welfare difference

$$W_a(\phi_A,\sigma) - W_{ab}(\phi_A,\sigma) = -\phi_B(c_{Ba}-c_{Bb}) + 2(1-\sigma)\phi_A\phi_B P.$$

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 7. We prove the statement for $j = a$. A symmetric argument applies

to the other case.

$$\begin{aligned} W_{ab}(\phi_A, 1) &\geq W_a(\phi_A, 0) \\ \iff -\phi_A c_{Aa} - \phi_B c_{Bb} &\geq 2\phi_A\phi_B U - \phi_A c_{Aa} - \phi_B c_{Ba} \\ \iff c_{Ba} - c_{Bb} &\geq 2\phi_A U \\ \iff d_B &\geq (2\phi_A U)/P. \end{aligned}$$

□