Information, Reputation and Ethnic Conflict

Dominic Rohner
University of Cambridge

November 19, 2006

Abstract

Empirical studies have found ethnic cleavages to play an important role in the occurrence of civil conflict. Surprisingly, theoretical research on ethnic conflict has been very scarce. In the present contribution a theoretical model of reputation and ethnic conflict is built. Depending on the information structure and the reputation cost of defecting, economic interaction can either result in (peaceful) trade or in appropriative conflict. Ethnic divisions affect the reputation cost of defection and therefore influence the conflict risk. It is shown what respective effects ethnic fractionalisation, polarisation and segregation have on the risk of conflict.

JEL Classification: C73, D74, F10, L14, Z13.
Keywords: Conflict, Ethnicity, Reputation, Information, Trade.

1 Introduction

Civil wars do not only bring fear and death to those people affected, they are also a major obstacle to economic growth and development[1]. It is therefore not surprising that in recent years considerable effort has been made in economics to explain why conflicts occur. An open and much debated question is whether civil wars can be explained by rebel leaders solving collective action and organisational problems or by existing grievances in the population. Most scholars would agree that often existing tensions, frustrations, inequalities, as well as a powerful instrumentation and canalisation of these grievances by rebel leaders, are needed for a civil war to occur. Even the most scrupulous leaders are not able to initiate popular resistance if no grievances at all are present. Small tensions and conflicted issues between groups, such as disputed economic

[1] *Address:* Faculty of Economics, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. dr296@cam.ac.uk.

Acknowledgements: I would like to thank Partha Dasgupta, Samuel Bowles, Robert Evans, Michael Findley, John Miller, David Myatt, Scott Page, Rajiv Sethi, Christopher Wallace, Diego Winkelried and Elizabeth Wood for their helpful comments. As well, useful discussions with conference and workshop participants in Amsterdam, Santa Fe NM and Columbus OH are gratefully acknowledged.

interactions on the individual level, can result in a full-blown civil war if they are widespread and skilfully manipulated by rebel leaders.

A striking feature of many civil wars in recent decades is that grievances and disputes often tend to occur along ethnic lines. In countries as diverse as Rwanda, Sudan, Guatemala or Angola, ethnicity has played a major role in the breakout of hostilities.

The present paper focuses on disputes between individual players over gains from economic interactions, which often lead to powerful grievances. More concretely, the source of conflict is disputed economic relations, and it will be assessed what roles information, reputation and ethnicity play in keeping economic relationships peaceful. Players peacefully sharing the economic gains from their interaction will be associated with an outcome of "(peaceful) trade", whereas the situation of players competing over economic benefits will be labelled "appropriative conflict". Cleavages between ethnic groups[2], and in particular factors such as ethnic fractionalisation, polarisation and segregation, will be shown to influence the reputation cost of disloyal business, and therefore to affect the likelihood of trade or conflict.

The concepts of ethnic fractionalisation, polarisation and segregation are used in the following way: A highly *fractionalised society* is defined as one with a great number of ethnic divisions and of distinct groups or tribes. Following Montalvo and Reynal-Querol (2005, p.797), a highly *polarised society* is defined as a society "where a large ethnic minority faces an ethnic majority". Polarisation is greatest if a society consists of two ethnic groups of equal size, and is smallest if a society consists of one homogenous group. *Ethnic segregation* is defined as the extent to which different groups in the society are kept separate. High (low) levels of segregation will correspond to a big (small) part of the players' time spent in intra-group interaction. More formal definitions of ethnic fractionalisation, polarisation and segregation are included in the sections three and four, where the main results concerning the impact of fractionalisation, polarisation and segregation on the likelihood of conflict are derived.

There is a growing literature about ethnicity and civil wars. The empirical evidence is not conclusive so far, and it is not only controversial whether ethnic cleavages matter, but also how they should be measured. Fearon and Laitin (2003) and Collier and Hoeffler (2004) have found that ethnic fractionalisation does not increase the risk of civil war outbreaks. By contrast, Vanhanen (1999), Sambanis (2001) and Collier, Hoeffler and Rohner (2006) conclude, using other data sources and measures, that ethnic fractionalisation increases the risk of civil wars. Cederman and Girardin (2006) explain the occurrence of conflict with ethno-nationalist exclusiveness.

A nonlinear impact of ethnicity on conflict could explain the contradictory findings of the empirical literature. Horowitz (1985) has performed a comparative analysis and has found that for fully homogenous and for fully heterogenous societies the risk of ethnic conflict is small, whereas the risk is greater for fewer

[2] The present contribution uses a very comprehensive definition of ethnicity, which applies to tribal, religious and linguistic groups.

big groups confronting each other. Using cross-country evidence, Collier and Hoeffler (1998) have come to a similar conclusion, namely that intermediate levels of fractionalisation are the most risky. Based on the results of these studies and their own estimations, Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005) argue that what drives ethnic conflict is not fractionalisation, but polarisation. As their evidence is convincing, it would be important to know for which (theoretical) reason polarisation could matter.

The issue of ethnic segregation has also received considerable attention in the literature. Some scholars have claimed that segregation increases the risk of ethnic conflict (Diez Medrano, 1994; Olzak, Shanahan, and McEneaney, 1996), while others have argued that segregation, taking the form of "partition", could in some cases actually be a solution to ethnic conflict (Horowitz, 1985)[3].

Surprisingly, theoretical studies building formal models of ethnicity and conflict have been scarce. Esteban and Ray (1999) develop a behavioral model of contest and bargaining between groups according to the distribution of certain characteristics. In more recent papers (2006a; 2006b), the same authors focus on ethnic mobilisation and rent-seeking and on the question why in a society with class and ethnic cleavages, the latter tend to be more salient than the former[4]. Another formal model addressing ethnic conflict has been built by Caselli and Coleman II (2006). They study the interaction between coalitions of different groups, whereas ethnicity increases the risk of conflict by enforcing coalition membership.

All these papers focus on aggregate players ("interest groups") rather than on individual players, and do not treat the respective effects of ethnic polarisation, fractionalisation and ethnic segregation.

The creation and impact of ethnic identities is another important topic in the literature about ethnic conflict (Basu, 2005; Sen, 2006).

Other related contributions are the ones by Fearon and Laitin (1996) and Tirole (1996)[5], which emphasise intra-group enforcement of group members' cooperation with players from other groups. It is shown that policing inside a group can ensure peaceful relations and a good collective reputation outside the group. However, while these scholars stress the existence of collective action issues inside a group that assure peace and a good collective reputation, they ignore the role of information, individual reputation concerns and various kinds of ethnic cleavages.

The literature on "liberal peace" is also relevant for the present paper (see Polachek, 1980; Oneal and Russett, 1999; Gartzke, Li, and Boehmer, 2001).

[3] Sambanis (2000) concludes, using cross-country evidence, that partition does not significantly prevent conflict occurrence.

[4] In another interesting paper Esteban and Ray (2006c) treat the contrasting effects of fractionalisation and polarisation on conflict onsets and intensity. However, they do not include ethnicity and reputation concerns in their model. Rather, fractionalisation, polarisation, as well as political institutions matter by affecting the cost of conflict for different groups in a collective action framework.

[5] The contribution of Tirole (1996) treats as well individual reputation concerns, but in a principal-agent model that emphasises reputation issues related to business and does not account for ethnic cleavages and conflict.

These scholars show how trade relationships between countries can reduce the scope of inter-state war by increasing the long-run gains from economic cooperation. The specification of my model is compatible with the findings of the literature mentioned above, as it treats trade and war as substitutes between which players have to choose, taking into account the reputation cost of defection.

The present contribution would like to address the shortcomings of the existing literature by building a theoretical model in which ethnicity matters through the channels of information and reputation, and that is able to assess the impact of ethnic polarisation, segregation and fractionalisation on the likelihood of ethnic conflict. Rather than focusing on coalition building and contests between aggregate groups, as has been done in previous studies, I will emphasize the potential for disputes and conflicts on the level of individual players. This has the advantage of avoiding *a priori* assumptions about the groups solving their collective action problems. An overall occurrence of a civil war will be seen as the sum of all individual-level disputes. This way of defining conflict seems reasonable as, in a society where all inter-group economic relations are dishonest and conflicted, the dangers of political unrest and civil conflict are imminent. Of course, the conflict potential and grievances rooted in individual level disputes are a necessary but not sufficient condition for conflict to occur. For a full-blown civil war to break out, individual grievances need to be instrumentalised or manipulated by rebel leaders and collective action has to be feasible. However, given that almost all theoretical papers on ethnic conflict focus on group mobilisation and collective action, it makes sense for the present contribution to concentrate on further ways in which ethnicity matters, i.e. by affecting the reputation cost of defection.

The theoretical framework of the present contribution will build on the existing literature on commitment, reputation and contract enforcement in trade and business (see Greif, 1993; Greif, Milgrom and Weingast, 1994; Tirole, 1996; Dixit, 2003).

The remainder of the paper will be organised as follows. Section 2 will be devoted to a basic model of peaceful versus conflicting economic interactions for a homogenous society. In section 3, ethnicity will be introduced in the model, and the impact of polarisation and segregation will be assessed. The model will be extended to n-groups in section 4 and the effects of fractionalisation will be studied. Section 5 concludes.

2 The Basic Model

In what follows I build a model of how reputation and information matter for determining whether the economic interaction between players is more likely to result in trade or in conflict. This basic model will provide us with a theoretical framework that allows us to analyse in which way ethnic cleavages affect the reputation cost of defection, and eventually the likelihood of conflict.

2.1 Strategies, Payoffs, Information

The following assumptions are made:

G.1 - General setting: The game lasts for an infinite number of periods. Players discount the future and take into account that, with some probability, they will "die" in a given future period. The players who die are replaced by newly born players. There is an infinite number of players who match randomly.

G.2 - Actions: All players have the choice between staying out, appropriating or trading. Engaging in trade or in appropriation are modelled as substitutes in the present framework. The relative scope of economic cooperation and conflict is captured by the variable F, which ranges from 0 (full trade) to 1 (full appropriation and conflict). We allow for intermediate values of F. However, it follows from the specification of the payoff function (below) that the variable F will always take the extreme values 0 (trade) or 1 (conflict). Under a trade regime the "cake" of economic benefits from interaction is peacefully split, whereas in the case of appropriative conflict the division of the "cake" is conflicted. The timing is as follows: First, players choose whether they enter into contact with the opponent, then they choose between engaging in appropriation and trading.

G.3 - Payoff function: In all periods all players receive a payoff of 0 if they stay out. If they enter into contact with their match, they have the payoff function displayed in equation (1)[6].

$$V_i = S(\frac{1}{2} + \theta(\rho F_i - \psi F_j)) - cF_i - gF_j \tag{1}$$

where i,j=players, S=economic gains (surplus) from interaction, θ=parameter capturing the decisiveness of fighting effort (with $0 \leq \theta \leq 0.5$), ρ=parameter indicating the fighting technology (ability) of player i ($0 \leq \rho \leq 1$), F=level of fighting effort ($0 \leq F \leq 1$), ψ=fighting technology of player j ($0 \leq \psi \leq 1$), c=parameter related to the cost of player i's fighting effort, and g=parameter measuring player i's cost inflicted by the fighting effort of player j.

The total economic gains of the interaction S are multiplied by the term $(\frac{1}{2} + \theta(\rho F_i - \psi F_j))$, which refers to the share that player i receives of the gains. In the "economics of conflict" literature, the expressions displaying the shares of a "cake" received by a particular player are called "contest success functions" (see Hirshleifer, 1989; Skaperdas, 1996). The term $(\frac{1}{2} + \theta(\rho F_i - \psi F_j))$ is a linear difference-form contest success function[7], where the relative share of player i depends linearly on the differences in fighting effort between the players. The shares of both players sum up to 1.

The parameter θ measures the decisiveness of the differences in the fighting effort between the two players. If $\theta = 0$, both opponents receive half of the "cake" S, independently of their fighting effort. By contrast, if $\theta = 1$, the level

[6] The payoff function of player j is analogous: $V_j = S(\frac{1}{2} + \theta(\psi F_j - \rho F_i)) - cF_j - gF_i$.

[7] The contest success function $(\frac{1}{2} + \theta(\rho F_i - \psi F_j))$ used for determining the size of the benefit shares is similar to the one used in Rohner (2006), although the present contribution introduces the fighting technology differently.

of fighting has a strong impact on the distribution of S. Further, the parameters ρ and ψ reflect the fighting technology of the two players. $\rho = 0$ indicates a total inefficient fighting technology of player i, where an increased fighting effort of i does not result in his obtaining a greater share. $\rho = 1$ corresponds to the case of a very efficient fighting technology of i. The case is analogous with player j's fighting technology ψ.

Player i's payoff function (1) also includes the parameters c and g which relate to the cost of his own fighting effort, respectively the destruction inflicted by the opponent's fighting effort.

G.4 - Types: There are two types of players who differ in their fighting ability ρ. The players referred to as "strong" ("weak") have $\rho = \alpha$ ($\rho = \beta$), where $\alpha > \beta$. A proportion p of the population are assumed to be "strong" types.

G.5 - Information: i) The players are incompletely informed about the type of the other players. All other features of the game such as the form of the payoff function, the strategy space and the distribution of the two types are common knowledge.

ii) In general, players only observe the actions played in the interactions in which they are involved. However, it is assumed that, if a player defects and his opponent cooperates, a proportion q of the players becomes informed about the defection. One could think, for example, of a player telling his friends about the bad behaviour of his last opponent. If both players defect, they do not inform their friends about the interaction. The intuitive reason is that they do not want to appear in a bad light, as they have defected themselves as well.

iii) The players are assumed to have imperfect recall. The players who learn in a given period about the cheating of another player will remember the fact that this player has defected in the past, without however remembering in which period(s) it happened. Also, players do not remember any other aspects of past interactions.

G.6 - Solution concept: The equilibrium concept used is the "Perfect Bayesian Equilibrium".

2.2 The equilibria of the stage game

First, I derive results which are valid for the stage game in any period, then I focus on the reputation cost of defection which is related to the "shadow of the future". Thus, for the moment we can think of the game as a one-shot game. It is analysed under what conditions players will enter into contact with their match, and whether they choose trading or appropriating.

In the case where players decide to enter the game, they choose appropriative activities ($F_i = 1$) rather than trade if $\rho > \rho^* = \frac{c}{S\theta}$. This cut-off level ρ^* is crucial for the equilibrium of the game.

Further, players only decide to enter the game if the payoff V_i (given the optimal levels of F_i and F_j chosen thereafter) is greater than their outside option of staying out, which equals 0.

If both types have very high fighting abilities, i.e. $\alpha > \rho^*$ and $\beta > \rho^*$, both will fight if they choose to enter. Equation (2) displays the condition under which players of a given type choose to enter. They enter if the expected payoff of entering is positive.

$$S(\frac{1}{2} + \theta(\rho - \widetilde{p}\alpha - (1-\widetilde{p})\beta)) - c - g > 0, \; where \; \rho \in \{\alpha, \beta\} \tag{2}$$

where $\widetilde{p} = \left(\frac{t_S p}{t_S p + t_W(1-p)}\right)$, p=proportion of the population being "strong" types, t_S=proportion of "strong" types entering, t_W=proportion of "weak" types entering.

The parameter $\widetilde{p}$ refers to the proportion of the entering players that are of a "strong" type. According to the values of the different parameters there are three possible outcomes: both types stay out, "strong" types enter and fight and "weak" types stay out or both types choose to enter and fight. For some ranges of values multiple equilibria arise. As the focus of the present contribution is the reputation effect of ethnicity, which is only relevant to the case of $\alpha > \rho^* > \beta$ treated further below, we will not go into more detail for the case of $\alpha > \rho^*$, $\beta > \rho^*$.

If both types have not very effective fighting technologies, $\alpha < \rho^*$, $\beta < \rho^*$, they will both choose full economic cooperation (the trade equilibrium), where $F_i = 0$. For both players trading, $F_i = F_j = 0$, the payoff of entering the game is always positive. Thus, for ineffective fighting technologies we end up in an equilibrium where all players choose the actions (enter, cooperate) in all periods.

The case which is most interesting and relevant to our research question is when $\alpha > \rho^* > \beta$. From now on we will focus on this case. For $\alpha > \rho^* > \beta$, "strong" types would in a one-shot game, if they enter, always choose appropriation ($F_i = 1$), and "weak" types would, if they do not stay out, always choose to trade ($F_i = 0$). As before, different cases can be distinguished according to the decision of the two types to enter or stay out. "Strong" types enter the interaction if condition (3) holds:

$$S(\frac{1}{2} + \theta(\alpha - \widetilde{p}\alpha)) - c - \widetilde{p}g > 0 \tag{3}$$

Please note that, for assuring correct and consistent beliefs, "strong" types must have the beliefs of all "strong" types entering if condition (3) holds. Thus, this implies that t_S=1 and condition (3) becomes: $S(\frac{1}{2}+\theta(\alpha-\widetilde{p}'\alpha))-c-\widetilde{p}'g > 0$, where $\widetilde{p}' = \left(\frac{p}{p+t_W(1-p)}\right)$.

"Weak" types enter the interaction if condition (4) holds:

$$S(\frac{1}{2} - \theta\widetilde{p}\alpha) - \widetilde{p}g > 0 \tag{4}$$

Given our assumption that $\alpha > \rho^* = \frac{c}{S\theta}$, the "strong" types have always greater incentives to enter the interaction than the "weak" types. To make the

analysis interesting, we can assume that condition (3) holds. It follows that "strong" types always enter, and "weak" types only enter the interaction with a given opponent if the probability $\widetilde{p}$ of the opponent being "strong" is smaller than some threshold level $\widetilde{p}^*$. Formally, this condition can be written as:

$$\widetilde{p} \leq \widetilde{p}^* = \frac{S}{2(S\theta\alpha + g)} \tag{5}$$

Again, for making the analysis interesting, we will assume that the proportion p of "strong" types is relatively small and that this condition holds if all "weak" players enter the game ($t_W = 1$). Accordingly, if condition (5) holds and if a "weak" type matches with some trader of whom she has not had any *a priori* information, she will enter the interaction and then choose to trade ($F_i = 0$). Please note that there is also another equilibrium where all "weak" players stay out, even though entering would be profitable if all "weak" types were to enter, and where accordingly $\widetilde{p} = 1$. For the rest of the analysis we will focus on the most interesting case where the conditions (3) and (5) hold, and where without *a priori* information "weak" players enter the game.

If, however, a player has learnt that her present opponent has "cheated" in the past, she can deduce (using Bayesian updating) that her opponent is with probability $\widetilde{p}$=1 a "strong" type and that therefore condition (5) does not hold (as far as $\widetilde{p}^*$<1, which we assume to be the case for making the analysis interesting). Thus, she will not enter the interaction.

2.3 The reputation cost of defection

So far, the stage game has been analysed. Now, inter-temporal considerations are included. As players both discount future benefits and take into account the possibility of dying in future periods, we multiply future benefits with multiples of the parameter $\delta = \delta_0 h$, where δ_0=discount factor ($0<\delta_0<1$) and h=probability of being still alive in a given period (0<h<1).

First of all, we have to state the infinite periods equivalent of equations (3) to (5), in order to know the conditions for which "strong" and "weak" types enter the game if they have not received any information about the past behaviour of the opponent (in the case of receiving information about a past cheating their belief structure is different, as we will see at a later stage). Taking into account the probabilities of opponents being of a "strong" type, and of defecting[8], conditional on having received no information, the equations (3) to (5) become:

$$S(\frac{1}{2} + \theta(\alpha - \widehat{p}\widehat{z}\alpha)) - c - \widehat{p}\widehat{z}g > 0 \tag{3'}$$

$$S(\frac{1}{2} - \theta\widehat{p}\widehat{z}\alpha) - \widehat{p}\widehat{z}g > 0 \tag{4'}$$

[8] As shown in the proof of Proposition 1, there exists no equilibrium where "weak" types defect.

$$\widehat{p} \leq \widehat{p}^{*} = \frac{S}{2(S\theta\alpha + g)\widehat{z}} \tag{5'}$$

where $\widehat{p}$=probability of the opponent being a "strong" type conditional on receiving no information about past defections, $\widehat{z}$=probability of the opponent defecting conditional on having not being informed about past defections.

We will focus on the case where the conditions (3') and (4') hold and where accordingly all "weak" players enter the game if they have no (negative) *a priori* information about their match. The optimal strategy of the "weak" types, and the optimal strategy of "strong" types who get informed about the opponent's past cheating are treated in proposition 1. Below, we will derive the optimal strategy of "strong types" when they receive no information about the past behaviour of their opponent. This is the aspect of our game that is most relevant for the study of ethnic conflict[9].

There is a reputation cost for "strong" types choosing appropriation in the first period, in the sense that the informed "weak" players will not enter in economic interaction with them. If this reputation cost is big enough, "strong" players will in the first period choose trade ($F_i = 0$) rather than conflict in order to avoid the reputation cost.

The condition under which "strong" players choose trade rather than conflict in the first period can be obtained by comparing the expected values of cooperation and defection. Usually such a problem would be very complex as one would have to consider an infinite number of strategies. Fortunately, the structure of the reputation cost of defection implies that "strong" types have only two potential strategies which are a best-reply for some parameter values: First, defection in the first period and always thereafter. Second, cooperation in all periods. In what follows it is shown that these two strategies are Perfect Bayesian equilibria for some parameter values. The preliminary results needed are derived in the lemmas 1 and 2.

As outlined earlier in the assumptions G.1 and G.5, at the beginning of each period a proportion of players die (some of which are informed about past defections), and then further people become informed about the defection in the past period. People who are informed, stay informed until their death. Allowing for some probability of "forgetting" would not affect our results.

[9] As shown in the proof of proposition 1, whenever a "strong" type is informed about a past defection of his opponent, he will choose conflict, knowing that his opponent will continue to defect. This dispute between two "strong" types does not result in anyone being informed, as both defect. This corresponds to the real-world example of rivalling gangs of criminals fighting each other clandestinely. As the public receives only little information, these disputes do not fuel ethnic conflict. By contrast, the information about cases of a "strong" type defecting on a "weak" type will be spread, and ethnic grievances can arise.

To show that "always defect" can be an equilibrium strategy for "strong" types for some parameter values, we have to show that "strong" types who find it in their interest to defect in some period (if the reputation cost of cheating is not big enough) will only continue to choose defection if the reputation cost of doing so is non-increasing, which is the case in our framework. The intuition of the proof is as follows:

A player will only defect in a given period if the initial gain of defection is greater than the loss due to the additional number of players informed about the cheating. At the beginning of the first period in which defection is chosen by a given player, nobody is informed about a past defection, as there was no past defection. As always, a part q of the "weak" non-informed players get informed[10]. After a second defection a proportion q of the non-informed "weak" players would become informed. As there are now less non-informed players (as some part of the informed players of the previous period survive), the additional reputation cost of defection in this second defection period would be smaller and so forth. Thus, once the player has defected, the reputation costs of future defections is ever decreasing.

Also, until the first defection the incentive structure of a given player is stationary. Thus, if he finds it in his interest to defect in a given period τ, he would already have incentives to start defecting at any period $t < \tau$. It follows that he will start defection in the first period.

The reasoning above is summarised in lemma 1.

Lemma 1 *A player who ever chooses to defect will start to do so in the first period, and will stick to defection in all future periods.*

Proof. Please refer to Appendix A. ■

The reasoning for the case of players choosing in all periods to "cooperate", treated in lemma 2, is similar to the reasoning applied to lemma 1. As long as a player never chooses defection, his incentive structure is the same for all periods and, if he does not find it in his interest to choose defection in a given period τ, he will not find it in his interest to do so in any future period $t > \tau$.

Lemma 2 *A player who finds it in her interest to play "cooperate" in a given period, will also choose "cooperate" in all previous and future periods.*

Proof. Please refer to Appendix A. ■

Following the results of lemmas 1 and 2, and assuming that the equations (3) and (5) hold, we can derive for "strong" types the conditions under which the equilibria "always cooperate" or "always defect" are selected when receiving a signal "N" (no information about the opponent's past behaviour is revealed). In all cases, "strong" players will choose conflict whenever they observe a signal

[10] Also the same proportion q of the already informed players get informed about the defection in that particular period, but this has no impact at all as they were already previously informed.

"I" (information that the opponent has defected in the past), as defecting on another defector will not result in a reputation cost (cf. assumption G.5).

Equation (6) represents the inter-temporal expected value for a given player i, who is of a "strong" type, to choose conflict in the first period and always thereafter. This expected value computation takes into account that in equilibrium opponents who have cheated in the past after observing "N" will cheat again, and that "weak" players who get informed in the future about player i's defection will stay out, while informed "strong" opponents will defect.

$$\begin{array}{l} \widehat{q}\left[\dfrac{S}{2}-c-g\right] \\ +(1-\widehat{q})\left\{\begin{array}{c} S(\frac{1}{2}+\theta\alpha(1-\widehat{z}\widehat{p}))-c-\widehat{z}\widehat{p}g \\ +\left(\frac{\delta}{1-\delta}-\widetilde{q}\right)\widehat{p}\left[S(\frac{1}{2}+\theta\alpha(1-\widehat{z}))-c-\widehat{z}g\right] \\ +\widetilde{q}\widehat{p}\left[\frac{S}{2}-c-\widehat{z}g\right]+\left(\frac{\delta}{1-\delta}-\widetilde{q}\right)(1-\widehat{p})\left[S(\frac{1}{2}+\theta\alpha)-c\right] \end{array}\right\} \end{array} \quad (6)$$

where $\widehat{q}$=expected probability of receiving a signal "I" (information that the present opponent has cheated in the past), $\widehat{z}$=expected proportion of potential "strong" type opponents who fight conditional on a signal of "N" (no information about the opponent's past behaviour is revealed), $\widehat{p}$=expected proportion of opponents being of a "strong" type conditional on having received a signal "N", $\widetilde{q}$=present value of the proportion of players who are informed in the different future periods about player i having cheated.

Equation (7) reports the expected value for "strong" types of choosing trade in all periods when receiving a signal "N", and choosing conflict when receiving a signal "I".

$$\widehat{q}\left[\frac{S}{2}-c-g\right]+(1-\widehat{q})\left\{\begin{array}{c} S(\frac{1}{2}-\theta\widehat{z}\widehat{p}\alpha)-\widehat{z}\widehat{p}g \\ +\frac{\delta}{1-\delta}\widehat{p}\left[S(\frac{1}{2}-\widehat{z}\theta\alpha)-\widehat{z}g\right]+\frac{\delta}{1-\delta}(1-\widehat{p})\left[\frac{S}{2}\right] \end{array}\right\} \quad (7)$$

The expected value of cooperation is greater if the number of players who are informed about the previous periods' defection(s) is big enough. "Strong" types will choose trade rather than conflict if condition (8) holds.

$$\widetilde{q}>\widetilde{q}^{*}=\frac{\frac{1}{1-\delta}(S\theta\alpha-c)}{\widehat{p}\left[S\theta\alpha(1-\widehat{z})\right]+(1-\widehat{p})\left[S(\frac{1}{2}+\theta\alpha)-c\right]} \quad (8)$$

Please note that the variable $\widetilde{q}$ is strictly increasing in q (this can be seen from the equations used in the proof of lemma 1). This permits us to focus in the following analysis on q (the cut-off level of q corresponding to $\widetilde{q}^{*}$ can be denoted as q*).

The equilibrium of the game is summarised in Proposition 1. The beliefs about the opponent cooperating are denoted by μ, where μ=probability that the opponent is a "strong" type.

Proposition 1 *The following set of strategies and beliefs constitutes a Perfect Baysian Equilibrium, if equations (3') and (4') hold and if* $\alpha > \rho^* > \beta$*:*

"Strong" types always choose (enter, defect; $\mu = 1$*) for a signal "I" and (enter, defect;* $\mu = \widehat{p}$*) for a signal "N". "Weak" types play (out;* $\mu = 1$*) if they observe a signal "I", and play (enter, cooperate;* $\mu = \widehat{p}$*) if they observe "N". This is an equilibrium if equation (8) does not hold, i.e. if* $\widetilde{q}$ *is small.*

"Strong" types play (enter, defect; $\mu = 1$*) for a signal "I" and (enter, cooperate;* $\mu = \widehat{p}$*) for a signal "N", "weak" types play (out;* $\mu = 1$*) after observing "I" and (enter, cooperate;* $\mu = \widehat{p}$*) after "N". This is an equilibrium if equation (8) holds, i.e. if* $\widetilde{q}$ *is big.*

This is the unique equilibrium for the "weak types" entering the game.

Proof. Please refer to Appendix A. ■

There are also two equilibria where the "weak" types do not enter the game. They are treated in the proof of Proposition 1. For the analysis of ethnic conflict in the next sections, only the case referred to in Proposition 1 is relevant.

In the next section it will be assessed how ethnic cleavages affect the (stage game) "reputation cost" of defection, q, and in this way influence the scope of trade and appropriative conflict.

3 Introducing Ethnicity in the Model

So far, the variable q has been regarded as exogenous. At present, q will become endogenous to the model and it will be discussed how ethnicity affects the level of q.

In a homogenous society with only one ethnic group, the probability of the next match of a player being informed about his past defection corresponds simply to the number of uninformed players ("friends"), that become informed by each player who has been betrayed in the previous period, divided by the total number of players in the population. Thus, $q = k$, where k=part of uninformed players that become informed about the defection. Please note that some of the players who become informed in a given period have already been informed. Thus, they do not matter for the analysis, which implies that we can exclusively focus on the uninformed players who become informed.

Introducing ethnicity in the model leads to additional assumptions and features of the game. These are listed below.

G.7 - Two groups: Initially, we assume that the population is composed of two groups, which differ in ethnic characteristics (in section 4 the model will be extended to n-groups). The first group amounts to a share w of the whole population ($0 \leq w \leq 1$). Accordingly, the part (1-w) of the population belongs to the second group.

G.8 - Part time d: Players spend a certain part d (whereas 0<d<1) of their total time endowment (which is normalized to 1) for within-group activities. For simplicity, it is assumed in the main text that this part d is fixed and does

not depend on the relative group sizes. This simplification allows the reader to more easily follow the derivations, and highlights the main results in a convenient way. However, the results are all robust for the consideration of a more general framework, with the part of time spent on intra-group interactions depending on the relative group sizes. Appendix B is devoted to the derivations and results of this more general case.

The model presented allows for different levels of d for different groups, and this general case will also be emphasised for the analysis of intra-group conflict. For inter-group conflict, however, we will generally assume without a loss of generality that all groups have the same level of d. This makes the equations more tractable and it becomes easier to understand what drives the main results. Allowing for different levels of d in the case of inter-group conflict does not alter the results.

The part d could for example be interpreted as the time spent on tribal gatherings or religious ceremonies and on other intra-group interaction. During this time, players only meet people from their own group. Similarly, the fraction of time (1-d) is spent with people from the other group. Typically, people from both groups are assumed to spend more than the proportional share of their time on intra-group activities. This can be expressed as $d>w$, $d>(1-w)$.

G.9 - Matching: For some values of d and w, not all players will find a trade partner. In this case they are assumed to get some compensation, for example through an insurance scheme set up by trade unions financed with a part of the gains from interaction. They are just outside the game for one period and will be in the same situation when they match with an opponent in the next period, i.e. the number of players informed about the defection remain the same as before. The situation of players who fail to find a partner and get compensated is different from the case where players choose not to enter and obtain the outside option, which is zero.

Fighting the opponent in a given period t implies that she will tell her friends. These to a large extent belong to the same ethnic group, and the proportion of people informed about the defection who are part of the same ethnic group as the opponent is larger than the proportion of informed players in other ethnic groups. If a given player fights an opponent from the same ethnic group as himself, many people whom the fighter meets will be informed about the defection. By contrast, fighting someone from another ethnic group will result in many people from this other ethnic group being informed about the defection. However, it is less likely to match with them in the next period. Intuitively, it becomes clear that people would in most cases have lower incentives to "cheat" on opponents from their own group and greater incentives to not be honest in relationships with players from other ethnic groups. Mathematically, the different levels of q for defecting on someone of one's own group or on an opponent from another group correspond to the weighted average of the conditional probabilities of being informed subject to being a member of a particular group times the likelihood of being met. The computation is done in Appendix A.

3.1 The Likelihood of Intra-Group and Inter-Group Conflict

The probability of the next period's match being informed about the present defection in the case of trade with a member of the same group is given by q_S below (computation in Appendix A).

$$q_S = k\left[\frac{d_i^2}{w} + \frac{(1-d_i)^2}{(1-w)}\right] \tag{9}$$

where, k=part of uninformed players who become informed about the defection ("friends"), w=relative size of the own group relative to the whole population ($0 \leq w \leq 1$), d_i=the part of the time a given player spends with people from her own group ($0 \leq d_i \leq 1$).

If a player defects on an opponent from another group, the probability q_D of the next period's match being informed becomes as displayed below. As mentioned earlier, we assume for convenience that $d_i = d_j$.

$$q_D = k\left[\frac{d(1-d)}{w(1-w)}\right] \tag{10}$$

Ethnic cleavages affect at the same time the unconditional probability of meeting people and the conditional probability of their being informed. The interaction of changes in these two values leads to non-linear effects of introducing ethnic cleavages on the likelihood of the next match being informed. In some cases, ethnic division can lead to increased conflict potential due to a lower reputation cost of fighting opponents. In other cases, the likelihood of conflict can be reduced. This is, for example, the case where the society has an extreme level of fractionalisation and segregation, resulting in a large number of totally autonomous communities with perfect monitoring.

It makes sense to start the formal analysis, as done in proposition 2, by deriving the conditions under which intra-group conflict is more or less likely than inter-group conflict.

Proposition 2 *The likelihood of intra-group conflict initiated by a member of a group i in ethnically divided societies is lower than the likelihood of inter-group conflict if the time spent for intra-group interaction is greater than group i's proportional share in the population ($d > w$), and if more than half of the total time available is spent on intra-group interaction ($d > 0.5$).*

Proof. Please refer to Appendix A. ■

Please note that both the relative magnitude of d_i (with respect to w), as well as the absolute value of d_i ($d_i > 0.5$) matter, for this condition to hold. Having assumed that people spend more time than corresponding to the proportional group share for intra-group interaction, we know that $d_i > w$, $d_i > (1-w)$.

Thus, the condition $d_i > 0.5$ of proposition 1 has to hold, as the more numerous group always has a population share greater than 0.5.

A related issue is, under which condition the level of intra-group conflict in ethnically divided societies is lower than the overall level of conflict in societies that are ethnically homogenous. The formal analysis of this issue in the Appendix leads to proposition 3.

Proposition 3 *The likelihood of intra-group conflict initiated by a member of a group i in religiously divided societies is lower than the likelihood of conflict in a homogenous society without cleavages, provided that the time spent on intra-group interaction is greater than the proportional share of group i in the population ($d_i > w$).*

Proof. Please refer to Appendix A. ■

The intuitive reason for this result is that enhanced intra-group interaction increases the reputation cost of defecting on a member of the same group. Therefore, peaceful trade becomes more likely.

As summarised in proposition 4, inter-group conflict in ethnically divided societies is greater than overall conflict in homogenous societies if $d + w > 1$. This is the case if our usual assumption of $d_i > w$, $d_i > (1 - w)$ holds.

Proposition 4 *The likelihood of inter-group conflict initiated by a member of group i in religiously divided societies is higher than the likelihood of conflict in a homogenous society without cleavages, provided that $d + w > 1$.*

Proof. Please refer to Appendix A. ■

3.2 The Impact of Polarisation

An important issue is how increases or decreases in polarisation affect the likelihood of intra-group and inter-group conflict. In the present paper I focus on the case of polarisation between two ethnic groups. As before, the population share of group i is given by w. The population share of the second group j is now labelled v. Both shares add up to 1 (w+v=1). Polarisation is defined in the following way.

Definition 1 *Polarisation*$=1 - |w - v|$*, where w=population share of group i, v=population share of group j.*

The more similar the shares of the two population groups, the higher is the level of polarisation in a given society.

The impact of changes in the population share of a group on its likelihood of intra-group cheating and conflict is given by the first derivative of q_S with respect w, as displayed in equation (11):

$$\frac{\partial q_S}{\partial w} = k\left[\frac{-d_i^2}{w^2} + \frac{(1-d_i)^2}{(1-w)^2}\right] \tag{11}$$

Reformulating, one can find that this expression is negative ($\frac{\partial q_S}{\partial w} < 0$), for $w < d_i$, i.e. when people spend more than the proportional share of their time on intra-group interaction. As discussed earlier, we can assume that this condition holds and that $\frac{\partial q_S}{\partial w} < 0$. In this case, increases in the size of his own group w (for a given level of group integration d) lead to more defection and thus a higher likelihood of intra-group conflict for player 1. An increase in w would however correspond to a decrease in v (as v=1-w), lowering the likelihood of intra-group conflict for the second player.

The impact of changes in w on the likelihood of inter-group conflict is given by expression (12).

$$\frac{\partial q_D}{\partial w} = k\left[\frac{d(1-d)}{(w(1-w))^2}\right](-1)\left[(1-w)-w\right] \tag{12}$$

The expression $\frac{\partial q_D}{\partial w}$ becomes positive for w>0.5.

Please note that the level of inter-group cheating is the same for both groups, independently of the parameter values, $q_D^1 = q_D^2$. Knowing that w+v=1, we can easily see from equation (10) that $q_D^1 = q_D^2$ always holds.

In the present framework, polarisation is defined as the relative strength of the two groups. Maximum polarisation corresponds to a case where both groups are of equal size, i.e. w=v=0.5. Minimum polarisation would correspond to a case where w=1, v=0 or w=0, v=1. This way of introducing polarisation in our theoretical framework is convenient, and consistent with the commonly used definitions and measures of polarisation (see Montalvo and Reynal-Querol, 2005).

It makes sense to analyse what happens if initially polarisation is a maximum (w=v=0.5) and if it is decreased afterwards by increasing the level of w (increasing v would lead to identical results). The effects on intra-group defection for the more numerous group (here group 1) and the less numerous group are summarised in proposition 5:

Proposition 5 *A marginal decrease in polarisation (increasing the population share of the more numerous group) for a given level of d_i results in a lower level of their q_S, provided that $d_i > w$, and accordingly in a higher level of intra-group conflict inside the more numerous group and in a lower level of intra-group conflict inside the smaller group.*

Proof. For d>w, in equation (10) we have $\frac{\partial q_S}{\partial w} < 0$. Thus, increasing (decreasing) w results in a lower (higher) q_S, and therefore a higher (lower) likelihood of defection and conflict. ■

As far as intra-group conflict is concerned, changes in levels of polarisation lead to less conflict in one group and more in the other group. It is a zero-sum-game, where what we gain on one hand, we lose on the other. As we have seen

in proposition 2, if some weak and reasonable assumptions hold, the likelihood of intra-group conflict is lower than the likelihood of inter-group conflict. In most countries, q_S is way above the critical level of q* (computed in equation (8)), and intra-group conflict never takes place.

In most countries that suffer from political instability and from ethnic conflicts, the constraint that is binding is the condition for inter-group conflict. As for d>w and d>0.5, which we can reasonably assume to always hold, $q_S > q_D$ and the likelihood of inter-group conflict is higher than for intra-group conflict, policy makers are mainly concerned about inter-group conflict.

The impact of reduced polarisation on inter-group conflict is summarized in proposition 6.

Proposition 6 *A marginal decrease in polarisation for a given level of d results in a higher level of q_D (for both groups) and accordingly in a lower level of inter-group conflict.*

Proof. For w>0.5 in equation (12) we have $\frac{\partial q_D}{\partial w} > 0$, and accordingly for w<0.5 we have $\frac{\partial q_D}{\partial w} < 0$. Thus, increasing w of the more numerous group results in a higher q_D (as w>0.5 $\Rightarrow \frac{\partial q_D}{\partial w} > 0$), whereas decreasing v of the smaller group results in a higher q_D as well (as v<0.5 $\Rightarrow \frac{\partial q_D}{\partial v} < 0$). ■

The present model framework provides a theoretical explanation as to why high levels of polarisation between ethnic groups can result in conflicts. This result has been found in the empirical literature (for example, Montalvo and Reynal-Querol, 2005), but theoretical models focusing on these issues have so far been sparse.

Figure 1 plots as a numerical example[11] the levels of q_S and q_D for different levels of w (group 1's population share). We can easily see that the values of q_D are lower than the values of q_S, indicating that the likelihood of inter-group conflict is higher than the likelihood of intra-group conflict. Thus, in most cases, reducing the scope for inter-group conflict becomes the main policy issue. The parameter q_D takes its lowest value at w=0.5 (maximum polarisation). It follows that the more polarised a society, the higher is the likelihood of inter-group conflict.

3.3 The Impact of Segregation

The concept of segregation refers to the separation and lack of interaction between different groups. The extent of segregation is measured in our model by the parameter d. The following definition applies:

Definition 2 *Segregation=d, where d=part of time spent for intra-group interaction.*

[11] The following parameter values have been used: d=0.8, k=0.25.

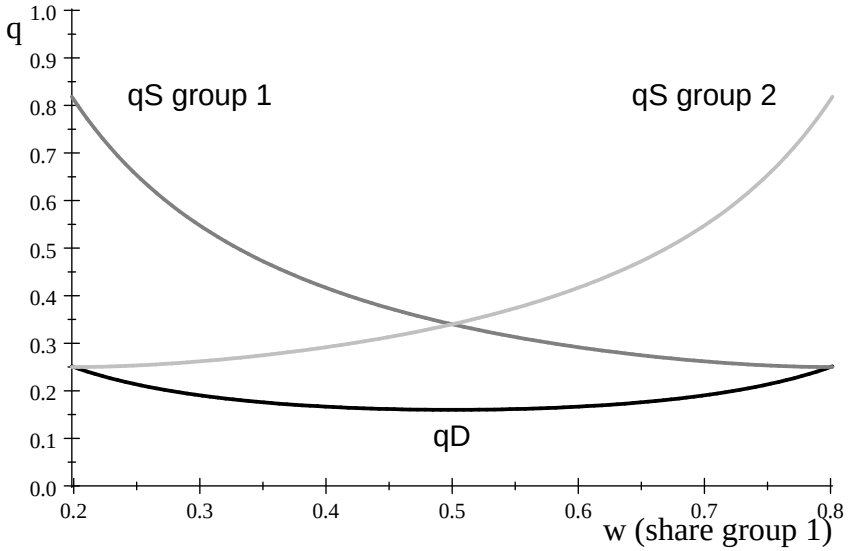

Figure 1: The Impact of Polarisation on qS and qD

High values of d correspond to a situation of strong segregation, with only little inter-group interaction. By contrast, low values of d indicate a very integrated society with a lot of inter-group interaction.

In history, it has often been an important issue as to whether segregation reduces or enhances conflict. Proponents of segregation policies have put forward that limiting the interaction between (hostile) population groups reduces the scope of conflict. For example, this logic has been followed for the construction of the Berlin Wall, the establishing of South African ghettos or, more recently, for the building of the wall separating Israel from the Palestinians. Opponents of such policies argue that separation walls create injustices and grievances that are the seeds for future conflicts.

Whether segregation policies are politically successful and morally justifiable is controversial. Neither the empirical nor the theoretical literature has been conclusive so far. The present theoretical framework aims to shed light on the effects of segregation on the likelihood of conflict. However, it does not provide normative or moral judgements about the appropriateness of such policies.

As for the case of polarisation, the analyse of segregation should first establish the impact of changes in the relevant parameter (which is d at present) on the likelihood of intra-group conflict. Equation (13) displays the first derivative of q_S with respect to d_i.

$$\frac{\partial q_S}{\partial d_i} = k \left[\frac{2d_i}{w} + \frac{-2(1-d_i)}{(1-w)} \right] \tag{13}$$

We have $\frac{\partial q_S}{\partial d} > 0$ if $d_i > w$. If we as usual assume the condition $d_i > w$ to hold, increases in d_i result in increases in q_S, and thus lead to reduced scope for intra-group conflict. This is summarised in the proposition below:

Proposition 7 *More segregation (a higher d_i) results in less intra-group conflict.*

Proof. From equation (13) follows $d_i > w \Leftrightarrow \frac{\partial q_S}{\partial d_i} > 0$. ■

This result is intuitive, as more intra-group interaction increases the possibility of the monitoring of intra-group cheating. Players have lower incentives to cheat on a trade partner from their own group, which reduces the likelihood of conflict.

However, as argued before, in most societies q_S is likely to be high and accordingly intra-group conflict is not very likely to occur. What is more often binding is the condition for intra-group conflict, as usually $q_D < q_S$. The impact of changes in d on q_D are given by equation (14) below.

$$\frac{\partial q_D}{\partial d} = k\frac{1}{w(1-w)}\left[1-2d\right] \tag{14}$$

It follows from equation (14) that $d > 0.5 \Leftrightarrow \frac{\partial q_D}{\partial d} < 0$. As discussed earlier, if the assumption d>w, d>(1-w) holds, it is required that d>0,5, as $\max\left[w, 1-w\right] > 0.5$. This being the case, segregation leads to a lower q_D, and thus to greater incentives for inter-group cheating and conflict.

Segregation increases the likelihood of inter-group conflict, but at the same time results in less inter-group interaction. For full segregation d=1, whenever inter-group relationships take place, the likelihood of defection will be very high. However, as d=1, no inter-group interaction actually takes place. Thus, full segregation leads to a lower overall conflict likelihood by reducing the likelihood of intra-group cheating. The decreased scope for defection is due to less inter-group interaction. Intuitively, in an extremely segregated world where people only interact in tiny villages with perfect monitoring, cheating would not take place.

For less than full levels of segregation, both the likelihood of inter-group interaction occurring and the likelihood of inter-group conflict conditional on inter-group interaction occurring should be taken into account. Segregation makes inter-group interaction less likely, but more likely if it takes place. For an initial situation where inter-group conflict occurs, $q_D < q^*$, segregation would reduce overall conflict by making inter-group interactions less likely. By contrast, if initially inter-group relations are honest and peaceful ($q_D > q^*$), increases in segregation bear a risk of increasing the overall likelihood of conflict, by decreasing q_D. For a substantial decrease in q_D, the condition $q_D < q^*$ might hold afterwards. Although at present less frequent, inter-group interaction might become conflicted, whereas initially it was peaceful. This reasoning is summarised in proposition 8.

Proposition 8 *Full segregation (d=1) eliminates inter-group conflict entirely. For intermediate levels of segregation (0<d<1), the impact of increases in segregation is ambiguous. For initially conflicted inter-group interaction ($q_D < q^*$), segregation reduces the occurrence of inter-group conflict. For initially honest and peaceful inter-group trade ($q_D > q^*$), segregation increases the scope for inter-group conflict.*

Proof. Follows from the reasoning discussed above. ■

4 Conflict in an n-group Framework

For analysing issues like polarisation it made sense to limit ourselves to a 2-group framework that allowed for an unequal size of the groups. However, for analysing fractionalisation, as well as for testing the robustness of previous results, it makes sense to use a n-group framework, with more than two groups each of an equal size. Fractionalisation is defined as below:

Definition 3 *Fractionalisation*$=1-\frac{1}{r}$*, where r=number of ethnic groups.*

The more ethnic groups there are, the higher the level of fractionalisation of a society. For only one ethnic group the level of fractionalisation is zero.

For intra-group defection, the likelihood q_S of the next period's opponent being informed about the dishonest behaviour is given by equation (15). Please note that q_S in the n-group framework corresponds to q_S in the 2-group framework with $w = \frac{1}{r}$, where r is the number of groups.

$$q_S = k\left[\frac{{d_i}^2}{1/r} + \frac{(1-d_i)^2}{(1-1/r)}\right] \tag{15}$$

The main difference between the n-group and the 2-group framework is that in the n-group case strangers from other groups do not all belong to the *same* other group. Thus, if a player from a group i cheats on an opponent of a given group j, this will result in a relatively high probability that other players of group j are informed of the cheating. However, players from another "foreign" group m will be as badly informed about the defection as the players of the "home" group i. Thus, it is necessary to take into account the probability of matching people from all different groups as well as their conditional probability of being informed. This has been done in Appendix A.

The likelihood of the next period's match being informed about the cheating, for inter-group conflict is given by equation (16).

$$q_D = k\frac{(1-d)r}{(r-1)}\left[2d + \frac{r-2}{r-1}(1-d)\right] \tag{16}$$

4.1 The Impact of Segregation and Fractionalisation in a n-group Framework

It is interesting to see whether propositions 7 and 8, summarising the effects of segregation on conflict, also hold in a n-group framework. This is the case if the first derivative of q_S with respect to d is positive, and the first derivative of q_D with respect to d is negative. As q_S is the same in the n-group framework as in the 2-group framework (for $w = \frac{1}{r}$), the results of the 2-group setting remain valid for n-groups. As far as q_D is concerned, its first derivative with respect to d is displayed in equation (17).

$$\frac{\partial q_D}{\partial d} = k\frac{r}{(r-1)}\left[-(2d + \frac{r-2}{r-1}(1-d)) + (1-d)(2 - \frac{r-2}{r-1})\right] \qquad (17)$$

We have $\frac{\partial q_D}{\partial d} < 0 \Leftrightarrow d > \frac{1}{r}$. Thus, the conclusions of proposition 8 in the previous section hold as well for the n-player framework, provided that the time spent for intra-group interaction, d, is greater than the proportional population share of each group, 1/r. As discussed earlier, this can generally be assumed to hold.

Fractionalisation is defined in the present framework as the number of (equally-sized) groups, r. For a given population size n, increasing the number of groups, r, would result in a greater number of smaller groups and in a society that is more fractionalised.

For assessing the impact of increases or decreases in fractionalisation on the likelihood of conflict, one can focus on the derivatives $\frac{\partial q_S}{\partial r}$ and $\frac{\partial q_D}{\partial r}$. For obtaining $\frac{\partial q_S}{\partial r}$, one can simply refer to the discussion of $\frac{\partial q_S}{\partial w}$ in the previous section. As $w = \frac{1}{r}$, $\frac{\partial q_S}{\partial r}$ has just the opposite sign as $\frac{\partial q_S}{\partial w}$ before. This result is summarised in proposition 9.

Proposition 9 *A marginal increase in fractionalisation (increasing the number of groups, r, in the population) for a given level of d results in a higher level of their q_S, provided that $d > \frac{1}{r}$, and accordingly in a lower level of intra-group conflict.*

Proof. Refer to the proof of proposition 5. ■

For assessing the impact of fractionalisation on inter-group conflict, we can take the derivative of q_D with respect to r.

$$\frac{\partial q_D}{\partial r} = kr(1-d)\left[\left(\frac{-1}{(r-1)^2}\right)(2d + \frac{r-2}{r-1}(1-d)) + \left(\frac{r}{r-1}\right)\left(\frac{1-d}{(r-1)^2}\right)\right] \qquad (18)$$

We have $\frac{\partial q_D}{\partial r} < 0 \Leftrightarrow d > \frac{1}{r}$. The interpretation of this condition is done in proposition 10.

Proposition 10 *If players spend more than the proportional time (according to the population share of a group) for intra-group interaction, $d > \frac{1}{r}$, fractionalisation increases the likelihood of inter-group conflict, by decreasing the reputation cost of defection.*

Proof. Follows from the discussion above. ■

This theoretical result is consistent with recent empirical evidence (cf., for example, Collier, Hoeffler and Rohner, 2006).

It is important to distinguish between segregation and fractionalisation. The intuitive explanation often made for a non-linear impact of fractionalisation on ethnic conflict is that, for high levels of fractionalisation, groups have lower incentives to behave cooperatively, but also have less opportunity for conflicts, as they interact less. What this argument is really about is not the impact of fractionalisation, but the potential conflict-reducing impact of complete segregation. Empirically, measures of segregation and fractionalisation can be correlated, as often rural and pre-industrial societies with a high level of fractionalisation would also have a high level of segregation, due to tedious and costly transports between the autonomous villages. This is the case for many traditional societies in Africa. If in such societies the observed level of conflict happens to be low, this is due to segregation rather than fractionalisation.

What is not captured by our model, and what could make the effects of fractionalisation on conflict more ambiguous, is the existence of collective action problems. As we have seen, fractionalisation increases the scope for inter-group conflict. However, it would be conceivable that fractionalisation makes collective action and the formation of a viable rebellion force composed by minorities more difficult. No single minority would have the critical mass to provide rebellion on its own, and coordinating the actions of different ethnic groups would be difficult. Thus, if fractionalisation increases grievances, but also increases the organisational costs of rebellion, its overall impact would be ambiguous, and a non-linear relationship between fractionalisation and conflict could emerge.

5 Conclusion

The present contribution has examined how ethnic divisions can result in either trade or conflict. The theoretical analysis has focused on individual-level interactions between players in a random matching framework. Defection has been associated to appropriative disputes, which can be a powerful source of grievances that are fuelling conflict. If conflicted interactions become widespread, the accumulation of initially minor economic disputes bears a significant risk of escalation and full-blown civil war.

What prevents players from defection in the present setting is the reputation cost of future opponents being informed about the dishonest behaviour. Ethnicity affects the reputation cost of defection by influencing the likelihood of matching in the future with a player who is informed about the cheating.

It has been shown in the paper that for weak and reasonable assumptions the likelihood of inter-group conflict is higher than of intra-group conflict, indicating that in most cases the binding constraint for achieving peace is to reduce inter-group conflict.

Increases in polarisation have been associated to a higher likelihood of inter-group conflict, and to more (less) intra-group conflict for the group that becomes more (less) numerous. Full segregation has been shown to reduce the scope for conflict. For intermediate levels of segregation, increases in segregation decreased the likelihood of intra- and inter-group conflict for initially conflicted interactions, and increased (decreased) the likelihood of inter-group (intra-group) conflict if interactions were initially peaceful. Fractionalisation has been found to increase the likelihood of inter-group conflict.

The present contribution has succeeded in building a unified theoretical framework for assessing the impact of polarisation, segregation and fractionalisation on ethnic conflict. It has been able to provide theoretical foundations for empirical results such as, for example, the conflict-enhancing impact of polarisation found by Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005) or the recent empirical evidence of fractionalisation increasing conflict (Collier, Hoeffler and Rohner, 2006). However, further research on related issues is much needed. Building theoretical models of ethnic conflict that account for collective action problems and identity construction and performing further empirical studies of the respective impact of polarisation, segregation and fractionalisation are strongly encouraged.

References

[1] Basu, Kaushik. (2005). "Racial conflict and the malignancy of identity", *Journal of Economic Inequality*, 3, 221-41.

[2] Caselli, Francesco, and Wilbur John Coleman II. (2006). "On the Theory of Ethnic Conflict", mimeo, London School of Economics and Duke University.

[3] Cederman, Lars Erik, and Luc Girardin. (2006). "Beyond Fractionalization: Mapping Ethnicity onto Nationalist Insurgencies", *American Political Science Review*, forthcoming.

[4] Collier, Paul and Anke Hoeffler. (1998). "On Economic Causes of Civil Wars", *Oxford Economic Papers*, 50, 563-73.

[5] Collier, Paul and Anke Hoeffler. (2004). "Greed and grievance in civil war", *Oxford Economic Papers*, 56, 563-95.

[6] Collier, Paul, Anke Hoeffler, and Dominic Rohner. (2006). "Beyond Greed and Grievance: Feasibility and Civil War", mimeo, University of Oxford and University of Cambridge.

[7] Diez Medrano, Juan. (1994). "The Effects of Ethnic Segregation and Ethnic Competition on Political Mobilization in the Basque Country, 1988", *American Sociological Review*, 59, 873-89.

[8] Dixit, Avinash. (2003). "Trade Expansion and Contract Enforcement", *Journal of Political Economy*, 111, 1293-1317.

[9] Esteban, Joan, and Debraj Ray. (1999). "Conflict and Distribution", *Journal of Economic Theory*, 87, 379-415.

[10] Esteban, Joan, and Debraj Ray. (2006a). "A Model of Ethnic Conflict", mimeo, Institut d'Anàlisi Económica and New York University.

[11] Esteban, Joan, and Debraj Ray. (2006b). "On the Salience of Ethnic Conflict", mimeo, Institut d'Anàlisi Económica and New York University.

[12] Esteban, Joan, and Debraj Ray. (2006c). "Polarization, Fractionalization and Conflict", mimeo, Institut d'Anàlisi Económica and New York University.

[13] Fearon, James, and David Laitin. (1996). "Explaining Interethnic Cooperation", *American Political Science Review*, 90, 715-35.

[14] Fearon, James, and David Laitin. (2003). "Ethnicity, Insurgency, and Civil War", *American Political Science Review*, 97, 75-90.

[15] Gartzke, Erik, Quan Li, and Charles Boehmer. (2001). "Investing in the Peace: Economic Interdependence and International Conflict", *International Organization*, 55, 391-438.

[16] Greif, Avner. (1993). "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition", *American Economic Review*, 83, 525-48.

[17] Greif, Avner, Paul Milgrom, and Barry Weingast. (1994). "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild", *Journal of Political Economy*, 102, 745-76.

[18] Hirshleifer, Jack. (1989). "Conflict and rent-seeking success functions: Ratio vs. difference models of relative success", *Public Choice*, 63, 101-12.

[19] Horowitz, Donald. (1985). *Ethnic Groups in Conflict*, Berkeley, University of California Press.

[20] Montalvo, José, and Marta Reynal-Querol. (2005). "Ethnic Polarization, Potential Conflict, and Civil Wars", *American Economic Review*, 95, 796-815.

[21] Olzak, Susan, Suzanne Shanahan and Elizabeth McEneaney. (1996). "Poverty, Segregation, and Race Riots: 1960 to 1993", *American Sociological Review*, 61, 590-613.

[22] Oneal, John, and Bruce Russett. (1999). "Assessing the Liberal Peace with Alternative Specifications: Trade Still Reduces Conflict", *Journal of Peace Research*, 36, 423-42.

[23] Polachek, Solomon. (1980). "Conflict and Trade", *Journal of Conflict Resolution*, 24, 55-78.

[24] Reynal-Querol, Marta. (2002). "Ethnicity, Political Systems, and Civil Wars", *Journal of Conflict Resolution*, 46, 29-54.

[25] Rohner, Dominic. (2006). "Beach holiday in Bali or East Timor? Why conflict can lead to under- and overexploitation of natural resources", *Economics Letters*, 92, 113-117.

[26] Sambanis, Nicholas. (2000). "Partition as a Solution to Ethnic War: An Empirical Critique of the Theoretical Literature", *World Politics*, 52, 437-83.

[27] Sambanis, Nicholas. (2001). "Do Ethnic and Nonethnic Civil Wars Have the Same Causes?", *Journal of Conflict Resolution*, 45, 259-82.

[28] Sen, Amartya. (2006). *Identity and Violence: The Illusion of Destiny*, New York, Norton.

[29] Skaperdas, Stergios. (1996). "Contest success functions", *Economic Theory*, 7, 283-90.

[30] Tirole, Jean. (1996). "A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)", *Review of Economic Studies*, 63, 1-22.

[31] Vanhanen, Tatu. (1999). "Domestic Ethnic Conflict and Ethnic Nepotism: A Comparative Analysis", *Journal of Peace Research*, 36, 55-73.

Appendix A - Derivations of the mathematical results of the main text

Proof of Lemma 1:

In each period first a proportion (1-h) of all players die and are replaced by newly born players, then a proportion q of players become informed if there was a defection in the previous period. Players remain informed until they die.

As for "strong" players, the condition $\alpha > \rho^* = \frac{c}{S\theta}$ holds, the initial gain of defection (labelled G_τ) in a given period τ equals $G_\tau = S\theta\alpha - c > 0$, which is the same in all periods.

The present value of the reputation cost L_τ of defecting for the first time in a given period τ equals the sum of all foregone gains in the future due to previously non-informed players getting informed about this particular defection. The exact cost of defection for a player i in a given period t depends on the actions chosen in the future. However, for any given strategy the cost of defection decreases in the number of past defections. Without loss of generality we can focus on the comparison between the reputation cost of defecting in period τ and defecting in period $\tau+1$ (the reasoning is the same for defecting in a period $t \geq \tau + 2$). It is found that $L_{\tau+1} = L_\tau(1 - qh) < L_\tau$, where q>0, h>0.

Below, this finding is illustrated by the derivation of the results for the two stationary cases of "always cooperate" and "always defect".

For the case of always playing (enter, defect) in all future periods $t \geq \tau + 1$, the loss of defecting in period τ (the first period of defection) equals $L_\tau = (\widetilde{q}_\tau^F - \widetilde{q}_\tau^T)y$, where $y = \widehat{p}\left[S\theta\alpha(1-\widehat{h})\right] + (1-\widehat{p})\left[S(\frac{1}{2}+\theta\alpha) - c\right]$ and where $\widetilde{q}_\tau^F = \left[q_\tau + \delta q_{\tau+1} + \delta^2 q_{\tau+2} + ...\right]$, with $q_\tau = 0$, $q_{\tau+1} = q$, $q_{\tau+2} = q_{\tau+1}h + (1 - q_{\tau+1}h)q = qh + (1-qh)q$ and so forth, and where $\widetilde{q}_\tau^T = \left[q_\tau + \delta q_{\tau+1} + \delta^2 q_{\tau+2} + ...\right]$, with $q_\tau = 0$, $q_{\tau+1} = 0$, $q_{\tau+2} = q$ and so forth. $\widetilde{q}_\tau^F$ ($\widetilde{q}_\tau^T$) refers to the present value of the proportion of players being informed in the future if fighting (trading) is chosen in period τ. The term capturing the reputation loss, $(\widetilde{q}_1^F - \widetilde{q}_1^T)$ of defecting for the first time in period 1, becomes $(\widetilde{q}_\tau^F - \widetilde{q}_\tau^T) = \left[\delta q + \delta^2(1-q)qh + ...\right]$.

If a player continues to defect in period $\tau + 1$ after having defected for the first time in period τ, the immediate gains G of defection remain the same, but the reputation cost is different. Without loss of generality we can treat the case of a defection in period $\tau+1$, after having already defected in period τ. The loss becomes $L_{\tau+1} = (\widetilde{q}_{\tau+1}^F - \widetilde{q}_{\tau+1}^T)y$, where $\widetilde{q}_{\tau+1}^F = \left[q_{\tau+1} + \delta q_{\tau+2} + \delta^2 q_{\tau+3} + ...\right]$, with $q_{\tau+1} = q$, $q_{\tau+2} = qh + (1-qh)q$ and $q_{\tau+3} = (qh + (1-qh)q)h + (1 - (qh + (1-qh)q)h)q$ etc, and $\widetilde{q}_{\tau+1}^T = \left[q_{\tau+1} + \delta q_{\tau+2} + \delta^2 q_{\tau+3} + ...\right]$, with $q_{\tau+1} = q$, $q_{\tau+2} = qh$ and $q'_{\tau+3} = qh^2 + (1-qh^2)q$ etc. It follows that $(\widetilde{q}_{\tau+1}^F - \widetilde{q}_{\tau+1}^T) = \left[\delta q(1-qh) + \delta^2(1-q)qh(1-qh) + ...\right] = (\widetilde{q}_\tau^F - \widetilde{q}_\tau^T)(1-qh)$. Given that q>0, h>0, we know that $(\widetilde{q}_{\tau+1}^F - \widetilde{q}_{\tau+1}^T)$ is smaller than $(\widetilde{q}_\tau^F - \widetilde{q}_\tau^T)$, and that defecting becomes less and less costly the more a player has defected in the past.

For the case of always playing (enter, cooperate) in all future periods $t \geq \tau + 1$, the gains of defection are as before, and the reputation cost of defecting in period τ (the first period of defection) equals $L_\tau = (\widetilde{q}_\tau^F - \widetilde{q}_\tau^T)y$, where $\widetilde{q}_\tau^F =$

$\left[q_\tau + \delta q_{\tau+1} + \delta^2 q_{\tau+2} + ...\right]$, with $q_\tau = 0$, $q_{\tau+1} = q$, $q_{\tau+2} = qh$ and so forth, and where $\widetilde{q}_\tau^T = 0$. It follows that $(\widetilde{q}_\tau^F - \widetilde{q}_\tau^T) = \widetilde{q}_\tau^F = \left[\delta q + \delta^2 qh + ...\right]$.

If a player defects in period $\tau + 1$ after having defected for the first time in period τ, and plays (enter, cooperate) in all future periods $t \geq \tau + 2$, the reputation cost of defection corresponds to $L_\tau = (\widetilde{q}_{\tau+1}^F - \widetilde{q}_{\tau+1}^T)y$, where $\widetilde{q}_{\tau+1}^F = \left[q_{\tau+1} + \delta q_{\tau+2} + \delta^2 q_{\tau+3} + ...\right]$, with $q_{\tau+1} = q$, $q_{\tau+2} = qh + (1 - qh)q$, $q_{\tau+3} = [qh + (1 - qh)q]\, h$ etc. Further, $\widetilde{q}_{\tau+1}^T = \left[q_{\tau+1} + \delta q_{\tau+2} + \delta^2 q_{\tau+3} + ...\right]$, with $q_{\tau+1} = q$, $q_{\tau+2} = qh$, $q_{\tau+3} = qh^2$ etc. Again, we obtain $(\widetilde{q}_{\tau+1}^F - \widetilde{q}_{\tau+1}^T) = \left[\delta q(1 - qh) + \delta^2 qh(1 - qh) + ...\right] = (1 - qh)(\widetilde{q}_\tau^F - \widetilde{q}_\tau^T)$.

The results obtained above for the stationary cases of always playing (enter, cooperate) or always playing (enter, defect) in the future, also hold for all non-stationary cases (if they were to exist), where players cooperate in some periods and defect in others.

To summarise, once a player chooses to defect in some period τ, the reputation cost of defection in any future period $t > \tau$ will be smaller than it was in period t. It follows that players who defect once will continue to defect in all future periods. Further, up to the period when the first defection occurs, the game is stationary and the incentives faced in each period are the same; therefore, if a player has incentives to first defect in a period τ, he would also have had incentives to defect in an earlier period $\tau' < \tau$. Thus, we have $\tau = 1$, i.e. the player defects in the first period.

Proof of Lemma 2:

A player will only choose "cooperation" in a given period τ if the present value of choosing "cooperate" is greater than of playing "defect". It is intuitive that if the reputation cost of defection is big enough, "strong" types would choose cooperation. Since in this case the stationary incentive structure would be the same for each period, if they are better off cooperating, it is in their interest to start with cooperation immediately in the first period.

Proof of Proposition 1:

First, we can treat all strategies where players choose the same actions independently of the signal they observe about the opponent.

1) Both "weak" and "strong" types always selecting (out; $\mu \in [0, 1]$) is an equilibrium, as no player would be better off deviating.

2) Both "weak" and "strong" types always choosing (enter, cooperate; $\mu \in [0, 1]$) is not an equilibrium, as "strong" types would deviate and play (enter, defect; $\mu \in [0, 1]$).

3) Both "weak" and "strong" types always selecting (enter, defect; $\mu \in [0, 1]$) is not an equilibrium, as that would not be the "weak" types' best reply.

Next, we can consider strategies when "weak" and "strong" types do both not condition their actions on their signals, but where different actions are chosen.

4) "Weak" types always choosing (out; $\mu \in [0, 1]$) and "strong" types always choosing (enter, cooperate; $\mu \in [0, 1]$) is not an equilibrium, as "strong" types would deviate and play (enter, defect; $\mu \in [0, 1]$).

5) "Weak" types always choosing (out; $\mu \in [0,1]$) and "strong" types always choosing (enter, defect; $\mu \in [0,1]$) is an equilibrium, as nobody would deviate.

6) "Weak" types always choosing (enter, cooperate; $\mu \in [0,1]$) and "strong" types always choosing (out; $\mu \in [0,1]$) is not an equilibrium, as "strong" types would deviate and play (enter, defect; $\mu \in [0,1]$).

7) "Weak" types always choosing (enter, cooperate; $\mu \in [0,1]$) and "strong" types always choosing (enter, defect; $\mu \in [0,1]$) is not an equilibrium, as "weak" types would deviate and play (out; $\mu = 1$) if they observe the signal "I".

8) Any cases where "weak" types always choose (enter, defect; $\mu \in [0,1]$) and "strong" types do not condition their actions on their signals cannot be equilibria, as "weak" types would deviate.

Now, we consider cases where at least one type conditions his actions on the signal observed. First, the cases are treated when "weak" do not condition their actions on the signal, while "strong" types do.

9) Any cases where "weak" types always select (out; $\mu \in [0,1]$) and "strong" types condition their actions on their signals cannot be equilibria, as the best reply of the "strong" types would be to always play (enter, defect; $\mu \in [0,1]$), as in the equilibrium found earlier.

10) Any cases where "weak" types always choose (enter, cooperate; $\mu \in [0,1]$) and "strong" types condition their actions on their signals cannot be equilibria, as the best reply of the "strong" types would be to always play (enter, defect; $\mu \in [0,1]$).

11) Any cases where "weak" types always choose (enter, defect; $\mu \in [0,1]$) and "strong" types condition their actions on their signals cannot be equilibria, as the best reply of the "strong" types would be to always play (enter, defect; $\mu \in [0,1]$).

At present, the cases of the "weak" type conditioning, and the "strong" type not conditioning are assessed.

12) Any cases where "strong" types always choose (out; $\mu \in [0,1]$) and "weak" types condition their actions on their signals cannot be equilibria, as the best reply of the "weak" types would be to always play (enter, cooperate; $\mu \in [0,1]$).

13) Any cases with "strong" types always choosing (enter, cooperate; $\mu \in [0,1]$) and "weak" types conditioning their actions on their signals cannot be equilibria, as the best reply of the "weak" types would be to always play (enter, cooperate; $\mu \in [0,1]$).

14) When "strong" types choose (enter, defect; $\mu = 1$) for a signal "I" and (enter, defect; $\mu = \widehat{p}$) for a signal "N", and when "weak" types play (out; $\mu = 1$) if they observe signal "I", and play (enter, cooperate; $\mu = \widehat{p}$) if they observe "N", it is an equilibria for equation (8) not holding, i.e. if $\widetilde{q}$ is small. In this case nobody has incentives to deviate and the beliefs are consistent. If equation (8) holds, this would not be an equilibrium, as "strong" types would be better off to play (enter, defect; $\mu = 1$) for a signal "I" and (enter, cooperate; $\mu = \widehat{p}$) for a signal "N".

15) Another case is when "strong" types always select (enter, defect; $\mu \in [0,1]$) and "weak" types play (out; $\mu \in [0,1]$) if they observe signal "I", and

play (enter, defect; $\mu \in [0,1]$) if they observe "N", this is not an equilibrium, as "weak" types would deviate.

16) When "strong" types always choose (enter, defect; $\mu \in [0,1]$) and "weak" types play (enter, cooperate; $\mu \in [0,1]$) if they observe signal "I", and play (out; $\mu \in [0,1]$) if they observe "N", this is not an equilibrium, as "weak" types would deviate.

17) Another case is when "strong" types always choose (enter, defect; $\mu \in [0,1]$) and "weak" types play (enter, cooperate; $\mu \in [0,1]$) if they observe signal "I", and play (enter, defect; $\mu \in [0,1]$) if they observe "N". This could not be an equilibrium, as an "innocent" "weak" player i knows that his opponent will always defect, as she will receive a signal "N". The only reason for player i to defect is to be rewarded by a "weak" player choosing (enter, cooperate; $\mu \in [0,1]$) in a future interaction. However, as his current match defects with certainty, a defection of player i would not result in anyone being informed. Thus, he is better off choosing (enter, cooperate; $\mu \in [0,1]$) or (out; $\mu \in [0,1]$) according to the parameter values.

18) A further case is when "strong" types always choose (enter, defect; $\mu \in [0,1]$) and "weak" types play (enter, defect; $\mu \in [0,1]$) if they observe signal "I", and play (out; $\mu \in [0,1]$) if they observe "N". A signal "I" can only come from a "strong" type who has defected on "weak" types. Thus, "weak" types would deviate, and it is not an equilibrium.

Now, the case is treated when both players do not condition their actions on their signal.

19) First, consider the case when "weak" types play (out; $\mu = 1$) after observing "I" and (enter, cooperate; $\mu = \widehat{p}$) after "N". There is an equilibrium when "strong" types play (enter, defect; $\mu = 1$) for a signal "I" and (enter, cooperate; $\mu = \widehat{p}$) for a signal "N" if equation (8) holds. The beliefs are consistent, and nobody has incentives to deviate. The case of equation (8) not holding is treated under 14).

20) Consider the case when "weak" types play (out; $\mu \in [0,1]$) after observing "I" and (enter, defect; $\mu \in [0,1]$) after "N". This is not an equilibrium without conditioning, as the "strong" types' best reply would be to always play (enter, defect; $\mu \in [0,1]$).

21) Consider the case when "weak" types play (enter, cooperate; $\mu \in [0,1]$) after observing "I" and (out; $\mu \in [0,1]$) after "N". However, this is not an equilibrium without conditioning, as the "strong" types' best reply would be to always play (enter, defect; $\mu \in [0,1]$).

22) Further, there is the case when "weak" types play (enter, cooperate; $\mu \in [0,1]$) after observing "I" and (enter, defect; $\mu \in [0,1]$) after "N". However, this is not an equilibrium without conditioning, as the "strong" types' best response would be to always play (enter, defect; $\mu \in [0,1]$).

23) When "weak" types play (enter, defect; $\mu \in [0,1]$) after observing "I" and (out; $\mu \in [0,1]$) after "N", this is not an equilibrium without conditioning, as the "strong" types' best reply would be to always play (enter, defect; $\mu \in [0,1]$).

24) Moreover, there is the case when "weak" types play (enter, defect; $\mu \in [0,1]$) after observing "I" and (enter, cooperate; $\mu \in [0,1]$) after "N". The best

reply of "strong" types would be to either always play (enter, defect; $\mu \in [0,1]$) or to play (enter, defect; $\mu \in [0,1]$) after observing "I" and (enter, cooperate; $\mu \in [0,1]$) after "N", according to the parameter values. In either case "weak" types would be better off deviating after observing "I".

Computing the probability of the next period's match being informed about the defection:

For intra-group defection, the overall probability q_S of the next match being informed is given by equation (A.1).

$$q_S = P(S)P(k \mid S) + P(D)P(k \mid D) \tag{A.1}$$

where, $P(S)$=Probability of meeting a player belonging to the same group, $P(k \mid S)$=Probability of the match being informed, conditional on being from the same group, $P(D)$=Probability of meeting a player belonging to another group, $P(k \mid D)$=Probability of the match being informed, conditional on being from another group.

By definition, the probability of a match in the next period with a player from one's own group is: $P(S) = d_i$, whereas d_i=the part of the time a given player spends with people from her own group ($0 \leq d_i \leq 1$). Accordingly, the probability of matching with someone outside the group becomes $P(D) = (1 - d_i)$.

Further, we have

$$P(k \mid S) = \frac{d_i}{w}k \tag{A.2}$$

where, k=part of uninformed players who become informed about the defection ("friends"), w=relative size of the own group relative to the whole population ($0 \leq w \leq 1$).

$$P(k \mid D) = \frac{(1 - d_i)}{(1 - w)}k \tag{A.3}$$

Introducing (A.2) and (A.3), together with $P(S) = d_i$, and $P(D) = (1 - d_i)$, in (A.1), we obtain:

$$q_S = d_i \left[\frac{d_i}{w}k\right] + (1 - d_i)\left[\frac{(1 - d_i)}{(1 - w)}k\right] = k\left[\frac{{d_i}^2}{w} + \frac{(1 - d_i)^2}{(1 - w)}\right] \tag{A.4}$$

For inter-group defection, the overall probability, q_D, of the next match being informed is again given by equation (A.1). As before, the likelihood of matching with a player of one's own group equals $P(S) = d$ (for convenience, it is assumed that for the case of inter-group conflict $d_i = d_j$). The probability of matching in a given period with someone of the group of last period's betrayed

opponent is $P(D) = (1-d)$. However, for the inter-group case, the conditional probability $P(k \mid S)$ becomes as displayed in equation (A.5). It is assumed that also in other groups the part d of their time is used for intra-group activities.

$$P(k \mid S) = \frac{(1-d)}{w}k \tag{A.5}$$

The conditional probability that defecting on a stranger will be known in the next period by a member of the stranger's group becomes:

$$P(k \mid D) = \frac{d}{(1-w)}k \tag{A.6}$$

Introducing $P(S)$, $P(D)$, (A.5), and (A.6) in (A.1), we obtain (A.7)

$$q_D = d\left[\frac{(1-d)}{w}k\right] + (1-d)\left[\frac{d}{(1-w)}k\right] = d(1-d)k\left[\frac{1}{w(1-w)}\right] \tag{A.7}$$

Proof of Proposition 2:

It can easily be seen from equation (8) that a higher probability of the next match being informed, q_i, reduces the likelihood of defection and increases the likelihood of peaceful trade. It follows that the likelihood of intra-group conflict is lower than the likelihood of inter-group conflict if $q_S > q_D$. We have $q_S = k\left[\frac{d^2}{w} + \frac{(1-d)^2}{(1-w)}\right]$ and $q_D = k\left[\frac{d(1-d)}{w(1-w)}\right]$. Proposition 2 is valid if proposition (A.8) holds.

$$q_S > q_D \Leftrightarrow \frac{d^2}{w} + \frac{(1-d)^2}{(1-w)} > \frac{d(1-d)}{w(1-w)} \tag{A.8}$$

Condition (A.8) holds if $d > w$ and $d > 0.5$.

Proof of Proposition 3:

The likelihood of intra-group conflict is lower than the general likelihood of conflict in homogenous societies, if $q_S > q$. We have $q = k$ and $q_S = k\left[\frac{d^2}{w} + \frac{(1-d)^2}{(1-w)}\right]$. Setting $k\left[\frac{d^2}{w} + \frac{(1-d)^2}{(1-w)}\right] > k$, we obtain after reformulation condition (A.9), which always holds.

$$q_S > q \Leftrightarrow (d-w)^2 > 0 \tag{A.9}$$

Proof of Proposition 4:

The likelihood of inter-group conflict $q_D = k\left[\frac{d(1-d)}{w(1-w)}\right]$ is higher than the overall conflict likelihood in a homogenous society q, if condition (A.10) holds.

$$q_D < q \Leftrightarrow \left[\frac{d(1-d)}{w(1-w)}\right] < 1 \tag{A.10}$$

This condition holds if $d + w > 1$.

Computing q_D for n-groups:

For inter-group cheating, the overall probability, q_D, of the next match being informed is given by equation (A.11).

$$q_D = P(S)P(k \mid S) + P(C)P(k \mid C) + P(T)P(k \mid T) \tag{A.11}$$

where, $P(S)$=Probability of meeting a player belonging to the same group, $P(k \mid S)$=Probability of the match being informed, conditional on being from the same group, $P(C)$=Probability of meeting a player belonging to the group of the present opponent, $P(k \mid C)$=Probability of the match being informed, conditional on being from the group of the present opponent, $P(T)$=Probability of meeting a player belonging to some third group, $P(k \mid T)$=Probability of the match being informed, conditional on being from some third group.

As before, the likelihood of matching with a player of one's own group equals $P(S) = d$. The probability of matching in a given period with someone of the group of last period's betrayed opponent is $P(C) = (1-d)\frac{1}{r-1}$, where r=number of groups. Further, $P(T) = (1-d)\frac{r-2}{r-1}$.

The conditional probabilities are as follows:

$$P(k \mid S) = P(k \mid T) = \frac{(1-d)\frac{1}{r-1}}{\frac{1}{r}}k \tag{A.12}$$

$$P(k \mid C) = \frac{d}{\frac{1}{r}}k \tag{A.13}$$

Introducing $P(S)$, $P(C)$, $P(T)$, (A.12), and (A.13) in (A.11), we obtain (A.14).

$$\begin{aligned} q_D &= d\left[\frac{(1-d)\frac{1}{r-1}}{\frac{1}{r}}k\right] + (1-d)\frac{1}{r-1}\left[\frac{d}{\frac{1}{r}}k\right] + (1-d)\frac{r-2}{r-1}\left[\frac{(1-d)\frac{1}{r-1}}{\frac{1}{r}}k\right] \\ &= k\frac{(1-d)r}{(r-1)}\left[2d + \frac{r-2}{r-1}(1-d)\right] \end{aligned} \tag{A.14}$$

Appendix B - Derivations for the extent of intra-group interaction depending on group sizes

In Appendix B the computations of sections 3 and 4 are re-done for the case where the part of time spent for intra-group interaction is not constant, but depends on the group size. The results of sections 3 and 4 are robust for this extension.

It is at present assumed that the players spend some fixed amount of time on intra-group and inter-group interaction, and that another part of their time is attributed to intra- or inter-group interaction depending on the relative group sizes. As before, we will first focus on the 2-group case. The new probabilities P(S) and P(D) are displayed below.

$$P(S) = d_i + (1 - d_i - e_i)w \tag{B.1}$$

$$P(D) = e_i + (1 - d_i - e_i)(1 - w) \tag{B.2}$$

As previously, it holds that $P(S) = 1 - P(D)$. Since players tend to spend more time with other players belonging to their own group, it is assumed that $d_i \gg e_i$ and that $P(S) > w$, which is the case for small or intermediate values of w and for the condition $d_i \gg e_i$ being fulfilled.

Computing the probability of the next period's match being informed of the defection:

As before, *for intra-group defection*, the overall probability q_S of the next match being informed is given by equation (B.3).

$$q = P(S)P(k \mid S) + P(D)P(k \mid D) \tag{B.3}$$

where, $P(S)$=Probability of meeting a player belonging to the same group, $P(k \mid S)$=Probability of the match being informed, conditional on being from the same group, $P(D)$=Probability of meeting a player belonging to another group, $P(k \mid D)$=Probability of the match being informed, conditional on being from another group.

Given (B.1) and (B.2), the conditional probabilities become:

$$P(k \mid S) = \frac{d_i + (1 - d_i - e_i)w}{w}k \tag{B.4}$$

where k=percentage of uninformed players who become informed about the defection ("friends"), w=relative size of the player's own group relative to the whole population ($0 \leq w \leq 1$).

$$P(k \mid D) = \frac{e_i + (1 - d_i - e_i)(1 - w)}{(1 - w)}k \tag{B.5}$$

Introducing (B.1), (B.2), (B.4) and (B.5) in (B.3), and after reformulation we obtain:

$$q_S = k\left[\frac{d_i^2}{w} + \frac{e_i^2}{1-w} + 1 - (d_i + e_i)^2\right] \tag{B.6}$$

For inter-group defection, the overall probability, q_D, of the next match being informed is again computed according to the same formula as in (B.3), and the values of $P(S)$ and $P(D)$ are the same as before (see (B.1), respectively (B.2)). The new conditional probabilities are displayed in the equations (B.7) and (B.8).

$$P(k \mid S) = \frac{e_j + (1 - d_j - e_j)(1-w)}{w}k \tag{B.7}$$

$$P(k \mid D) = \frac{d_j + (1 - d_j - e_j)w}{(1-w)}k \tag{B.8}$$

Introducing (B.1), (B.2), (B.7) and (B.8) into (B.3), we obtain after reformulation (B.9).

$$q_D = k\left[\frac{d_i e_j}{w} + \frac{d_j e_i}{1-w} + 1 - (d_i + e_i)(d_j + e_j)\right] \tag{B.9}$$

Propositions 2, 3 and 4:

The probability $P(S)$ can be expressed as $P(S) = d_i + (1 - d_i - e_i)w = D$. It follows that $P(D) = 1 - D$, $q_S = k\left[\frac{D^2}{w} + \frac{(1-D)^2}{(1-w)}\right]$ and $q_D = k\left[\frac{D(1-D)}{w(1-w)}\right]$. The formula of q_S and q_D , expressed in terms of D are exactly equivalent to the formula of (9) and (10) expressed in terms of d. Thus, the results of propositions 2 to 4 also hold for the new specification.

Polarisation:

For assessing the impact of changes in polarisation on intra-group defection, we have to take the first derivative of q_S with respect to w.

$$\frac{\partial q_S}{\partial w} = k\left[\frac{-d_i^2}{w^2} + \frac{e_i^2}{(1-w)^2}\right] = k\left[\frac{e_i^2 w^2 - d_i^2(1-w)^2}{w^2(1-w)^2}\right] < 0 \tag{B.10}$$

This derivative is negative, as before in the main text, if e is relatively small, and w not too large. That is the case for small and intermediate values of w and if our initial assumption of people spending a more than proportional part of their time on intra-group interaction holds (i.e. $d_i \gg e_i$). Thus, as before our model predicts that less polarisation leads to more (less) intra-group conflict inside the group that sees its population share increase (decrease).

Also, as far as the results for inter-group conflict are concerned, the conclusions of the main text are robust to letting $P(S)$ vary for changes in the

population size (for better highlighting the effects of polarisation, we set $d_i = d_j, e_i = e_j$):

$$\frac{\partial q_D}{\partial w} = k\left[\frac{-de}{w^2} + \frac{de}{(1-w)^2}\right] = k\left[\frac{de(2w-1)}{w^2(1-w)^2}\right] \gtrless 0 \Leftrightarrow w \gtrless 0.5 \tag{B.11}$$

As in the main text, the derivative is positive for w>0.5 and negative for w<0.5, indicating that decreases in polarisation (making the population shares of the two groups less equal) result in less conflict.

Segregation:

In the model of the main text we had $P(S) = d_i$, and more segregation simply corresponded to an increase in d_i. At present, $P(S) = d_i + (1 - d_i - e_i)w$, and again increased segregation is represented by a greater (fixed) part of time spent on intra-group interaction, d_i. The impact of segregation on intra-group conflict is displayed below in equation (B.12).

$$\frac{\partial q_S}{\partial d} = k\left[\frac{2d_i}{w} - 2(d_i + e_i)\right] = 2k\left[d_i\left(\frac{1}{w} - 1\right) - e_i\right] > 0 \tag{B.12}$$

The derivative $\frac{\partial q_S}{\partial d}$ is positive if e_i is relatively small compared to d_i, $d_i \gg e_i$, and if w is not too large. Again, this confirms the main text's previous results of segregation reducing intra-group conflict.

The results for inter-group conflict are as follows (again for simplicity we set $d_i = d_j, e_i = e_j$):

$$\frac{\partial q_D}{\partial d} = k\left[e\left(\frac{1}{w} + \frac{1}{1-w}\right) - 2(d+e)\right] < 0 \tag{B.13}$$

We obtain a negative derivative, as in the main text, if $d \gg e$, indicating that segregation makes inter-group interaction more conflicted, although less frequent.

Comparative statics of q_S for n-groups:

As shown previously in the main text, for intra-group conflict the analysis of segregation in a n-group framework is identical to a two-group framework, as the relevant equations are congruent for $w = \frac{1}{r}$.

Again, as before in the main text, the effect of fractionalisation on the likelihood of intra-group conflict, $\frac{\partial q_S}{\partial r}$, is simply the inverse of the effect of $\frac{\partial q_S}{\partial w}$ computed in equation (B.10).

Comparative statics of q_D for n-groups:

As before, for inter-group defection, the overall probability, q_D, of the next match being informed is given by equation (B.14).

$$q_D = P(S)P(k \mid S) + P(C)P(k \mid C) + P(T)P(k \mid T) \tag{B.14}$$

where, $P(S)$=Probability of meeting a player belonging to the same group, $P(k \mid S)$=Probability of the match being informed, conditional on being from the same group, $P(C)$=Probability of meeting a player belonging to the group of the present opponent, $P(k \mid C)$=Probability of the match being informed, conditional on being from the group of the present opponent, $P(T)$=Probability of meeting a player belonging to some third group, $P(k \mid T)$=Probability of the match being informed, conditional on being from some third group.

The relevant expressions become (for simplicity we consider $d_i = d_j, e_i = e_j$):

$$P(S) = d + (1 - d - e)\frac{1}{r} \tag{B.15}$$

$$P(C) = \frac{e}{r-1} + (1 - d - e)\frac{1}{r} \tag{B.16}$$

$$P(T) = e\left[\frac{r-2}{r-1}\right] + (1 - d - e)\frac{r-2}{r} \tag{B.17}$$

$$P(k \mid S) = P(k \mid T) = \frac{\frac{e}{r-1} + (1 - d - e)\frac{1}{r}}{\frac{1}{r}} \tag{B.18}$$

$$P(k \mid C) = \frac{d + (1 - d - e)\frac{1}{r}}{\frac{1}{r}} \tag{B.19}$$

Introducing equations (B.15) to (B.19) in (B.14), we obtain after reformulation:

$$q_D = k\left[\frac{2der}{r-1} + \frac{e^2(r-2)r}{(r-1)^2} + 1 - (d+e)^2\right] \tag{B.20}$$

For assessing the impact of fractionalisation on inter-group conflict we take the first derivative of q_D with respect to r.

$$\frac{\partial q_D}{\partial r} = 2ek\left[\frac{e - d(r-1)}{(r-1)^3}\right] < 0 \tag{B.21}$$

The derivative $\frac{\partial q_D}{\partial r}$ becomes negative for $d > e$ and at least two groups $r \geq 2$. This is consistent with the previous result in the main text that fractionalisation (a higher r) makes inter-group interactions more conflicted.

The impact of changes in segregation is analysed below:

$$\frac{\partial q_D}{\partial d} = 2k\left[e\left(\frac{r}{r-1} - 1\right) - d\right] < 0 \tag{B.22}$$

As in the main text, we have $\frac{\partial q_D}{\partial d} < 0$ for $d > e$ and at least two groups $r \geq 2$, which implies that segregation makes the (less frequent) inter-group interactions more conflicted.