Social Sanctions in Interethnic Relations:

The Benefit of Punishing your Friends

Christian Stoff*

March 2005

Abstract: We analyse interethnic cooperation in an infinitely repeated prisoner's dilemma when members of one group are unable to target punishment towards individual defectors from the other group. We first show that indiscriminate outgroup punishment may sustain cooperation in this setting. Our main result, however, is that the introduction of ingroup punishment in addition to outgroup punishment represents a better sanctioning institution in the sense that cooperative outcomes may persist in situations where outgroup punishment alone fails to induce cooperation. Our findings are consistent with historical evidence on the dynamics of interethnic conflicts.

Keywords: Interethnic conflicts, Interethnic cooperation, Ethnicity, Intergroup relations, Ingroup punishment, Outgroup punishment

JEL Classification: C70, C72, O12, O17, Z13

*Address: University of Zurich, Socioeconomic Institute, Bluemlisalpstr. 10, 8006 Zurich; Phone: +41.1.634 3996, e-mail: cstoff@soi.unizh.ch

I am grateful to my thesis advisor, Armin Schmutzler, for his support. I am particularly indebted to Michael Kosfeld for most helpful discussions. I would also like to thank Ernst Fehr and Urs Fischbacher as well as the seminar audience at the University of Zurich for their suggestions.

1 Introduction

Over the last twenty years, the nature of conflicts has changed dramatically. Whereas the typical conflict for which the United Nations has been set up was between nations, we observe that more recent conflicts are not necessarily across borders but rather between ethnic groups.[1]

These interethnic conflicts are clearly multi-faceted. Each conflict is different in its details and causes. A full understanding demands an interdisciplinary approach on a case by case basis. The ethnological literature offers a large number of highly valuable case studies of interethnic relations which form the basis on which other social scientists may build their theories. However, they also provide a variety of different theories on the roots and causes of interethnic conflicts. These theories can be classified into two branches. The first branch involves a social psychological approach viewing ethnic groups as firmly bounded, durable communities with a strong sense of a group identity. Within this branch, ethnic conflicts are the product of a deeply rooted ethnocentrism with a natural hostility towards outsiders. The motives to engage in interethnic conflicts are based on passion and involve a strong identification with the group.[2] The second branch can be viewed as a rationalist approach where ethnic groups are perceived as social constructs sharing a solidarity towards material rewards. Here, ethnic conflict are based on calculation. As soon as the material rewards of an ethnic group are threatened, the members of the group are willing to engage in interethnic conflicts.[3]

In this model, we will take a different approach that uses an economist's

[1] During the romanian-american symposium on inter-ethnic relations in Bucharest in 1991, Donald Horowitz already pointed out that the concept of the nation-state is an obsolete one, having lost both of its original meanings–the melting pot and the ethnic state.

[2] A branch of social psychology analyses intergroup relations in the experimental laboratory (see, for example, Brewer, 1979). They identified a preference for ingroup members even under a minimal group setting where players were randomly allocated to each group. However, it remains an unsolved question whether the origin of this ingroup bias lies in the stronger identification with the ingroup or whether it is based on calculative grounds such as generalized reciprocity motives (Yamagishi and Kyonari, 2000).

[3] A synthesis of both the passionate and the calculative theories is advocated by Horowitz (1998).

perspective. We do not attempt to develop a full theory of interethnic conflicts but we will concentrate on the role of information in interethnic relations. We will analyse the informational problems involved in interethnic interactions and their consequences for maintaining cooperation between ethnic groups. We will show that even with perfectly rational agents, interethnic conflicts arise from time to time. The following particularity seems to bear some commonality in the outbreak of interethnic conflicts. Whatever the origin of the hostility may be, one small incidence where some arbitary member of one group harms a member of the other group often triggers a widespread aggression where all players of both groups get involved. More specifically, individuals from the group of the harmed person retaliate against randomly chosen individuals from the group of the initiator which indicates that some form of a collective punishment is applied across groups. This is particular because both, the punishing individuals as well as the punished individuals are usually not directly affected from the incidence between the two individuals causing the escalation. There exist numerous examples from history which resemble this common pattern of collective outgroup punishment (Greif et al., 1994, Hasluck, 1954, Boehm, 1994, Dumont, 1982).[4] Similar stories accumulate in today's newspapers, suggesting that there is an urge for finding proper political responses to these problems. In this paper, we attempt to tackle two main questions. First, how can such a collective punishment towards another group be explained within a model with perfectly rational agents? Second, given that this dangerous form of collective outgroup punishment is chosen by ethnic groups, are there ways to prevent these escalations by introducing supplementary countervailing institutions?

To answer these questions, we set up the following game-theoretic model: We consider a model with two ethnic groups with two players in each group.[5] In each period, one player from each of the two groups is randomly matched in an outgroup pairing where they play a Prisoner's Dilemma with the choice

[4] An overview is provided by Fearon and Laitin (1996).

[5] In our model, ethnic group boundaries are exogenous. We want to focus on the consequences of a population structured into ethnic groups. See McElreath et al. (2003) on the evolution of ethnic markers in a population playing a coordination game. Further, see Bowles and Gintis (2004) for the evolution of group boundaries who emphasize the role of information in social interactions with incomplete contracts.

between cooperation and defection. This stage game is infinitely repeated and the strategies are common knowledge. We assume that the root of the problem in ethnical conflicts is informational. Although we assume that all players are able to observe the outcomes of the outgroup pairings, we make the assumption that across groups, individuals cannot identify other individuals. This entails that in each outgroup pairing, no player from either group knows with whom she is matched. The only thing they know is that they are matched with some player from the other group. This is in contrast to the informational setting within groups where identification of individual group members is possible, that is, each player knows the personal histories of their group members. We are interested in finding punishment strategies that induce the players in the outgroup pairings to cooperate despite this informational problem across groups.

In our model, we propose two punishment strategy profiles. The first punishment strategy profile involves an outgroup punishment strategy and has all players from both groups withholding cooperation in all subsequent outgroup pairings that follow as soon as an arbitrary player of an arbitrary group defects. Given our identification assumption, players from either group are not able to discipline individual players from the other group by threatening to punish them directly for their potential defection. However, since we assume that each player always has a positive probability of being chosen for an outgroup pairing, a collective punishment towards the other group which involves defecting against any player from the other group in the subsequent outgroup pairings following the first defection may still serve as disciplining device for potential defectors. We show that for a given exogenous parameter regime, this punishment strategy is a subgame perfect Nash equilibrium (SPNE) with cooperation in the outgroup pairings on the equilibrium path. The punishment that is applied off the equilibrium path is of a collective nature and may be referred to as an interethnic conflict. This allows us first to explain why unaffected individuals defect in periods following an incidence where their group member met a deviant player from the other group. Second, it becomes clear why these individuals punish some random person from the other group. In our model, the collective form of punishment across groups is a direct response to the assumed informational problem.

Despite its ability to induce cooperation on the equilibrium path, the equilibrium is however unappealing. Off equilibrium, following the first incidence of defection of an arbitrary player of an arbitrary group, all players from both groups that are matched in the subsequent outgroup pairings regardless of their personal histories of actions are punished. More figuratively, innocent players are being torn into an ethnic conflict by a single defection of an arbitrary player of an arbitrary group.

The second punishment strategy profile deals with this inefficiency off the equilibrium path. Here we introduce the possibility that players may directly punish their ingroup member that has been matched in the outgroup pairing. The enforcement of the ingroup punishment is assumed to be costly for both the punishing player and the player that is punished. The incentives to punish an ingroup member for her defection are given by the threat of a collective outgroup punishment phase. In the combined strategy profile, the outgroup punishment phase is initiated whenever a player defects in an outgroup pairing without being punished sufficiently by her group member.

A comparison of the two punishment strategies yields our main finding. Even if there is no cooperation with an outgroup punishment strategy profile for a range of exogenous parameters, the combined strategy profile is still able to achieve cooperation in our model. More specifically, a combined punishment institution which uses the above mixture of outgroup and ingroup punishment enables cooperative equilibria for a wider range of exogenous parameters than the institution with outgroup punishment alone. This will be true in particular if the cost of imposing the required sanction is less than the loss to the punished player that is required in the combined strategy profile.

Our model bears relation to several strands of literature.[6] We receive our empirical backing from numerous ethnological case studies from both anthropologists and historians.[7] The wide range of ethnological theories offers different approaches which should be viewed as complementary to our model.

[6] We will only present a short review of the related literature here. A more detailed discussion of the relevant literature is provided in the third part of this paper.

[7] See, for example, Cohen (1969), Unesco (1974), Greif et al. (1994), Dumont (1982), Hasluck (1954).

The experiments of social psychologists offer interesting insights on the psychological factors involved in ingroup and outgroup relations.[8] Methodogically, we use a game-theoretic approach to interethnic conflicts where we resort to the recent developments of network effects within groups. These network effects allow players to discipline players within a group to cooperate even if the players do not have long-term relationships with the same partner but there is random matching within the group instead. In particular, we use the insights of the community enforcement mechanism put forward by Kandori (1992). Our model is perhaps closest to Fearon and Laitin (1996) who analyse an institution with in-group policing which uses ingroup punishment alone and a spiral regime which uses outgroup punishment alone. In their model, they analyse the two institutions separately. However, we believe that there are additional mechanisms from the interaction between ingroup and outgroup punishment that are lost in such a separate analysis. We provide a simultaneous treatment and we highlight these interactions effects. In their model, the ingroup punishment does not involve any costs for the punisher. We believe that the introduction of an ingroup punishment that is costly to the punished player as well as the punisher is more realistic. In our argument, we do not fall back on indirect reciprocity or altruistic motives (Fehr and Fischbacher, 2004, Falk et al., 2001) although these motives are certainly involved in interethnic relations as well. We keep the assumption of perfectly rational players instead and we show that even then, players punish their ingroup members despite the costs involved.

We view the outbreak of ethnic conflicts as a consequence of an informational problem that inhibits any conditioning of strategies on personal histories of actions of individual players from the other group. This assumption can be motivated as follows. Ethnic groups are often characterised by dense social networks which allow them to easily exchange information on past behaviour of other group members. One may imagine that especially

[8] See Bornstein (2003) for an approach that analyzes interethnic conflicts as team games between groups. His work deals with the dilemma originating in the differences between individual, group and the collective interest which are inherent in interethnic conflict situations. See Yamagishi and Kyonari (2000) on the motives for ingroup favoritism based on generalized reciprocity. For a review on the ingroup bias in minimal intergroup situations, we refer to Brewer (1979).

small ethnic groups have a well-developed system of gossip and rumours which allows to get information about other group members at very low costs.[9] Interactions across ethnic groups frequently lack these information transmission possibilities. Whereas players know the personal histories of their group members, they are not able to attach any corresponding labels onto the players from the other group. Fearon and Laitin (1996) recognize this informational asymmetry between and within groups in their model as well and emphasize in the introduction that "any institutional regime for maintaining cooperation across groups must somehow address the problem this asymmetry poses" (p.719).

The consequence of this lack of information assumption about the players from the other group is that no player is able to learn about the history of play of the player they are currently matched with. This includes that even if the players in the current outgroup pairing have already been matched in preceding outgroup pairings, they are not able to recognize each other. Our assumption is not designed to capture interethnic interactions between players (e.g. two traders) that interact in long-term relationships with each other on a frequent basis. Instead, we concentrate on the class of interethnic interactions which are characterised by informational asymmetries and relatively infrequent outgroup pairings between the same two agents.

A better understanding of the mechanisms involved in interethnic relations not only helps to better understand history but it provides valuable insights for the development of trust-enhancing institutions which were and still are necessary to enable exchange markets to function properly. One important historical lesson in the establishment of exchange markets in developing countries is that these countries lack the necessary trust-enhancing pre-state institutions which were present when state institutions were introduced in Western societies in the late medieval period. The missing pre-state institutions are among the reasons why the introduction of state institutions did not yield the same functioning markets in the developing world as we ex-

[9] Fearon and Laitin (1996) mention the role of institutions such as churches, schools or respected elders which may be understood as delimiters of ethnic groups by certifying and advertising the reputations of individual players within their ethnic community. See also Colson (1974) who emphasizes the network effect within ethnic groups through gossip.

pected from modern western societies. We can learn from historical analysis how the necessary decentralized institutions were formed in medieval trade and these insights could be helpful in developing countries today.

The paper is structured as follows. Part 2.1 presents the social dilemma between the groups. Part 2.2.1 introduces an institution with outgroup punishment alone which solves the social dilemma for a exogenous parameter regime to be specified. However, it is based on a punishment institution that is not satisfactory off the equilibrium path. One incidence of a defection triggers a collective punishment between both groups comparable to an interethnic conflict. Part 2.2.2 tries to find an institution which mitigates this problem through the supplementary introduction of an ingroup punishment yielding a combined punishment institution. Part 2.3 compares the two punishment institutions and analyses in which cases the combined punishment strategy profile yields more cooperation. Finally, in part 3, we look at the related literature and in part 4, we provide a conclusion.

2 The model

2.1 Basic setup

Imagine two ethnic groups $K = A, B$. In each ethnic group, there are two players K_i with $i = 1, 2$.[10] They play an infinitely repeated stage game Γ_F which runs as follows.

Matching Process. In each period, an arbitrary player from one group is matched with an arbitrary player from the other group. Such a match will subsequently be referred to as an outgroup pairing. We assume that there is only one outgroup pairing per period.

Definition 1 *An* ***outgroup pairing*** *has a player A_i randomly matched with a player B_j from the other group.*

In each period, one arbitrary player from each group is chosen for the outgroup pairing. The matching in each period t is independent. Let $\mu(A_i, t)$ be

[10] See the concluding section for a motivation of why we consider only two players in each group.

player $A_i^{\prime}s$ match at time t where the set can be possibly empty. We assume that player i is randomly drawn for an outgroup pairing in period t from group K according to a uniform distribution. Since the sample space has only two elements[11] drawn from a uniform distribution, the probability for player A_i to be matched with any player from the other group B is

$$\Pr\left\{\cup_{j=1}^{2}(\mu(A_i,t)=B_j)\right\} = \Pr\{A_i\}\cdot\left[\sum_{j=1}^{2}\Pr\{B_j\}\right] = 1/2\cdot 1 = 1/2 \equiv \eta_o.$$

so that η_o denotes the probability with which an arbitrary player from an arbitrary group K_i is drawn for an outgroup pairing in each period.[12]

Payoffs. Both players in the outgroup pairing play a prisoners' dilemma (PD). The players simultaneously decide about whether to cooperate $a_{K_i} = \{c\}$ or to defect $a_{K_i} = \{d\}$. The action space is therefore $a_{K_i} \in \{c, d\}$.

$$\text{Player } A_i \quad \begin{array}{c|c|c|} \multicolumn{1}{c}{} & \multicolumn{2}{c}{\text{Player } B_j} \\ \multicolumn{1}{c}{} & \multicolumn{1}{c}{\text{c}} & \multicolumn{1}{c}{\text{d}} \\ \cline{2-3} \text{c} & \Pi^{cc}, \Pi^{cc} & \Pi^{cd}, \Pi^{dc} \\ \cline{2-3} \text{d} & \Pi^{dc}, \Pi^{cd} & \Pi^{dd}, \Pi^{dd} \\ \cline{2-3} \end{array} \tag{1}$$

The normal form representation of the static game is as in Matrix (1). By the PD game structure, payoffs are such that $\Pi^{cc} < \Pi^{dc}$, $\Pi^{cd} < \Pi^{dd}$ and $\Pi^{cc} > \Pi^{dd}$. In a one-shot situation, choosing d is the dominant strategy for both players in any outgroup pairing. The equilibrium outcome (d, d) is Pareto inefficient compared to the outcome where both players cooperate, i.e. (c, c).

Information structure. The strategies of the players are assumed to be common knowledge. After the two players in the outgroup pairing interact, all players from both groups observe the outcome so that there is perfect information transmission among all players concerning the outcomes of the

[11] The sample space is $\{K_1, K_2\}$ for each group $K = A, B$, i.e. either player K_1 or player K_2 from each group is chosen for the outgroup pairing.

[12] The probability for player A_i to meet a specific player B_j is

$$\Pr\{\mu(A_i,t)=B_j\} = \Pr\{A_i\}\cdot\Pr\{B_j\} = 1/4 \qquad \text{for } i,j=1,2.$$

outgroup pairings.[13] However, there is an idenfication problem across groups.

Assumption 1 *Although all players are able to observe what has been chosen in the outgroup pairings, a player A_i is not able to tell which particular player B_j from the other group was responsible for the action choice. A player K_i can only identify and know the personal history of action choices of an individual player K_j from her own group, with $i \neq j$.*

This assumption essentially means that all players of an ethnic group look alike for players of the other ethnic group. The only characteristic which is observable across groups is the affiliation of the player to the other group. Only players of the same ethnic group can distinguish between individual players in their group.[14]

Note that, with this assumption, it is not possible to develop punishment strategies that condition on the past history of the individual players from the other group. Although it is known to the players that there was an incidence of a defection by a player from the other group, it is not possible for players in our model to recognize this player in future outgroup pairings. Individual punishment of defectors from the other group is therefore not applicable as a means to induce cooperation across groups. We have to find other strategies in order to achieve cooperative outcomes.

2.2 Social sanctions analysis

2.2.1 Outgroup punishment only

In this section, we analyse how cooperation in outgroup pairings might be achieved through sanctions across groups with an identification problem between the groups. We assume that there is no state who enforces sanctions

[13] More realistic models could be developed where only the players involved in the outgroup pairings know of the outcomes. The transmission of this information to the other players in the groups would have to be modeled (Buskens and Weesie 1999). There might be the possibility of "telling lies" about the outcomes. However, we do not model these possibilities here since we want to focus on the effect of an identification problem across groups.

[14] We refer to the introduction for a motivation of this assumption. See also Fearon and Laitin (1996).

against defecting players of both groups. Further, the only punishment that players can apply across groups is to reciprocate defection by withholding cooperation on their part in subsequent outgroup pairings.[15]

Consider the following community enforcement mechanism. The inability to identify players across groups destroys the possibility for the group member of the victim or the victim herself to target a punishment on a player who previously chose defection in an outgroup pairing. The only possible punishment that players across ethnic groups can apply is a collective form of punishment.

Definition 2 *An* ***outgroup punishment (OGP)*** *has all players from one group, whenever chosen for an outgroup pairing, defect against all players from the other group with whom they are paired.*

Let $h_t = \{a^0, ..., a^{t-1}\}$ denote the history of action choices of the players in the outgroup pairings up to period $t-1$ where $a^\tau = (a^\tau_{A_i}, a^\tau_{B_j})$ denotes the vector of action choices by player A_i and B_j in period τ. Each history h_t $(t = 0, ..., \infty)$ satisfies one of the following conditions:

I. For $t = 0$, $h_t = \varnothing$.

The type I history refers to the initial period's history.

II. $\nexists a^\tau$ for $\tau \in \{0, ..., t-1\}$ such that $a^\tau_{K_j} = d$ for at least one K_j

The type II history refers to the class of subgames where no player defected in any of the previous outgroup pairings.

III. $\exists a^\tau$ for $\tau \in \{0, ..., t-1\}$ such that $a^\tau_{K_j} = d$ for at least one K_j

The type III history refers to the class of subgames where at least one player defected in any of the previous outgroup pairings.

Consider the following strategy θ_{K_i}.

[15] We do not consider other forms of retaliation across groups such as physically attacking defectors or denying them access to territories. Boyd and Richerson (1992) refer to these alternative forms of punishment as retribution strategies.

Definition 3 *Let the **outgroup punishment strategy** θ_{K_i} be defined as*

$$a_{K_i,t}(h_t) = \begin{cases} c & \textit{if } h_t \textit{ is of type I or II} \\ d & \textit{if } h_t \textit{ is of type III} \end{cases}.$$

In words, this breaks down to the following. In outgroup interactions, players start with cooperation. As soon as an arbitrary player of an arbitrary group defects, switch to defection against all players from the other group forever.[16]

Consider an infinite repetition of the stage game Γ_F and let δ denote the discount factor.[17] The strategy profile Θ, where both players play strategy θ_{K_i}, describes an equilibrium in which cooperation between groups can be achieved on the equilibrium path by the threat of an outgroup punishment for all subsequent periods as defined above. The following proposition derives the condition under which this strategy profile is a SPNE.

Proposition 1 *The strategy profile Θ, with all players in both groups playing strategy θ_{K_i}, is a SPNE if and only if*

$$\frac{\Pi^{dc} - \Pi^{cc}}{\Pi^{cc}} \leq \frac{\delta}{1-\delta}\eta_o. \tag{2}$$

The proof is given in *Appendix A.1.* Intuitively, the left-hand term $\frac{\Pi^{dc}-\Pi^{cc}}{\Pi^{cc}}$ represents the short-term deviation gain from defection over cooperation expressed in percentage terms. This short-term gain must be smaller than the right-hand side which represents the gain from having interethnic cooperation over an interethnic conflict in the form of defection in all subsequent outgroup pairings.

We have proposed one possible strategy profile that may induce intereth-

[16] The length of the outgroup punishment phase is not crucial in deriving our main argument. For the general case of a $T \in \{0, ..., \infty\}$ period outgroup punishment phase, we refer to Stoff (2004).

[17] A finite repetition of the game Γ_F does not yield cooperative outcomes on the equilibrium path. A backward induction argument makes this point clear. Consider the last period. Then the game essentially becomes a one-shot situation with the familiar outcome prediction of (d, d). The prediction of the last period outcome again destroys any incentives to cooperate in the next to last period and so on.

nic cooperation with the threat of outgroup punishment. However, it is not without "adverse effects". A player's action choice has a direct effect on the payoff of the player from the other group she is matched with in the outgroup pairing. However, in a world where identification of players from other groups is not possible and players choose collective forms of outgroup punishment as a consequence, there are indirect effects as well. The defector is not the only one to receive defection in all subsequent outgroup pairings. If a player defects, her action choice creates a huge negative externality on all the other player because she alone ignites a collective outgroup punishment that affects all subsequent outgroup pairings in our model.

More figuratively, we can also interpret such an outgroup punishment phase as an interethnic conflict. With our strategy profile from above, we have modelled the interethnic conflict as an off-equilibrium event. One incident of an arbitrary member of an ethnic community haggling someone from the other ethnic community is enough to set in motion a spiral of violence between the two groups.[18] Case studies indicate that collective forms of an outgroup punishment seem to be highly relevant from an empirical point of view.[19] We believe that these indirect effects pose one of the key problems for achieving interethnic cooperation.

However, as we will show in the next section, this is not the only possible strategy profile that is able to achieve interethnic cooperation. In the outgroup punishment strategy profile, the ingroup members of the players in the outgroup pairing remain completely passive. Despite their ability to identify a potential defector in their group, they do not apply any punishment on their group member for defecting against the player from the other group. However, they suffer from the defection of her group member through the outgroup punishment phase that follows the defection. In the next section, we will propose a punishment strategy profile which uses the identification possibility of players within each group by giving them the possibility of

[18] By this reason, the term spiral equilibrium is used to define the equilibrium found by Fearon and Laitin (1996) which also uses a trigger strategy with a collective punishment regime between groups.

[19] Greif (1997) gives examples of collective punishments from the Middle Ages. Hasluck (1954) provides a study on Albanian blood feuds. See also Boehm (1994) and Dumont (1982).

punishing their group member. It will combine outgroup and ingroup punishment to induce cooperation across ethnic groups.

2.2.2 Combined punishment

We introduce a new two-stage game Γ_{FP} which runs as follows. We add the possibility of an ingroup punishment. Let F_{K_i} denote an action choice that corresponds to a punishment imposed by a player of one group on another player of the same group.

In the first stage of Γ_{FP}, players in the outgroup pairing play the original game Γ_F. In the second stage of Γ_{FP}, each player K_i that has not been chosen for the outgroup pairing has to decide about how much ingroup punishment $F_{K_i} = [0, \infty)$ she imposes on her group member K_j that was chosen for the outgroup pairing in the first stage of Γ_{FP}.

We assume that, whenever an ingroup punishment is imposed, it is costly for both players, the player on whom it is imposed and the player who imposes the punishment. The higher the choice of F, the higher the loss incurred to the punished player and the higher the cost for the punishing player as well.[20] The following definition summarizes the ingroup punishment.

Definition 4 *An* ***ingroup punishment (IGP)*** *is a sanction F_{K_i} imposed on an individual player K_j by her group member K_i, $i \neq j$. This sanction involves a loss of $F^D_{K_j}$ for the sanctioned player where*

$$F^D_{K_j}(F_{K_i}) = F_{K_i}$$

and a loss of $F^P_{K_i}$ for the sanctioning player with

$$F^P_{K_i}(F_{K_i}) = f(F_{K_i}) \tag{3}$$

where $f : D \rightarrow \mathbb{R}_+$, $D \subseteq \mathbb{R}_+$, is a monotonically increasing function in D.

[20] These costs can be interpreted as the losses to the punisher from creating an unpleasant atmosphere within her group through the execution of an ingroup punishment. We exclude the possibility that a punisher enjoys punishing a group member.

Let $\underline{F^D} > 0$ denote the minimum loss to a punished player that is demanded in the specification of the strategies of each player in order to forgive the defection of the punished player in a preceding period. Let $\underline{F}$ denote the corresponding punishment choice of the punishing player.[21] Let $h_t = \{x^0, ..., x^{t-1}\}$ denote the history of action choices up to period $t-1$ where $x^\tau = (a^\tau, F^\tau)$ with $a^\tau = (a^\tau_{A_i}, a^\tau_{B_j})$ and $F^\tau = (F^\tau_{A_j}, F^\tau_{B_i})$ denotes the vector of action choices by the players A_i and B_j chosen for the outgroup pairing and the players A_j and B_i who decide about the ingroup punishment in period τ. Each history h_t $(t = 0, ..., \infty)$ satisfies one of the following conditions:

I. For $t = 0$, $h_t = \varnothing$.

The type *I* history refers to the initial period's history.

II. $\nexists x^\tau$ for $\tau \in \{0, ..., t-1\}$ such that $(a^\tau_{K_j} = d$ for at least one $K_j) \wedge (F^\tau_{K_i} < \underline{F}$ with $F^D_{K_j} < \underline{F^D})$

The type *II* history refers to the class of subgames where, in all preceding periods, both players either have cooperated or the defecting player has been punished by her group member in the second stage of the same period.[22]

III. $\exists x^\tau$ for $\tau \in \{0, ..., t-1\}$ such that $(a^\tau_{K_j} = d$ for at least one $K_j) \wedge (F^\tau_{K_i} < \underline{F}$ with $F^D_{K_j} < \underline{F^D})$

The type *III* history refers to the class of subgames where there exists a preceding period where a defecting player was not punished by her group member with at least $\underline{F}$ in that period.

In order to define a proper strategy $\phi_{K_i}(\underline{F^D})$ which prescribes not only an action choice for the players in the first stage but also for the players in

[21] In the strategies below, we assume that all players demand the same minimum punishment $\underline{F}$. This way we exclude the possibility that one group or player demands a higher minimum punishment to make her forgive a defection, i.e. that some players expect different ingroup punishments.

[22] Essentially, this means that the victim's group forgives the defection of a previous period when the defector is ingroup punished with a punished player's loss of $F^D \geq \underline{F^D}$ in that period.

the second stage of Γ_{FP}, i.e. the ingroup punishment decision $F_{K_i,t}$, we have to classify all the types of subgames which may arise up to the second stage of Γ_{FP}. Let $\widetilde{h_t}$ denote the history of each class of subgames. Fix a group K. Each history $\widetilde{h_t}$ $(t = 0, ..., \infty)$ satisfies one of the following conditions:

$\widetilde{I}$. Suppose h_t to be of type I or II; and $a^t_{K_j} = c$.

$\widetilde{II}$. Suppose h_t to be of type I or II; and $a^t_{K_j} = d$.

$\widetilde{III}$. Suppose h_t to be of type III; and an arbitrary $a^t_{K_j}$.

Consider the following combined punishment strategy $\phi_{K_i}(\underline{F^D})$.

Definition 5 *Let the* ***combined punishment strategy*** $\phi_{K_i}(\underline{F^D})$ *be defined as*

$$a_{K_i,t}(h_t) = \begin{cases} c & \textit{if } h_t \textit{ is of type } I \textit{ and } II \\ d & \textit{if } h_t \textit{ is of type } III \end{cases}$$

and

$$F_{K_i,t}(\widetilde{h_t}) = \begin{cases} 0 & \textit{if } \widetilde{h_t} \textit{ is of type } \widetilde{I} \textit{ or } \widetilde{III} \\ \underline{F} & \textit{if } \widetilde{h_t} \textit{ is of type } \widetilde{II} \end{cases}$$

for $i \neq j$.

In words, the strategy $\phi_{K_i}(\underline{F^D})$ breaks down to the following. In outgroup interactions, start with cooperation. If it is observed that an arbitrary player of an arbitrary group defects and this defection is punished by the defector's group member with $F^D \geq \underline{F^D}$, then continue with cooperation.[23]
However, if this defection is not punished by the defector's group member with $F^D \geq \underline{F^D}$, then defect in all subsequent outgroup pairings.[24]
In ingroup interactions, do not punish ingroup members when they cooperate. Do not punish them either when they defect in an outgroup punishment phase. However, as soon as your ingroup member chose defection in the first

[23] This means that the players of the other group forgive the defection when it is punished by the defector's group member.

[24] Note that ingroup punishment at some later period does not provoke cooperation of the players of the other group.

stage of Γ_{FP} during an outgroup cooperation phase, then choose to punish the defector by imposing a sanction of $F^D \geq \underline{F^D}$ on her.[25]

Consider an infinitely repeated play of the two-stage game Γ_{FP}.[26] The following proposition derives the conditions under which the combined punishment strategy profile $\Phi(\underline{F^D})$, with all players in both groups playing strategy $\phi_{K_i}(\underline{F^D})$, is a SPNE.

Proposition 2 *Suppose*

$$\overline{F^P} = \frac{\delta}{1-\delta}\eta_o \Pi^{cc} \tag{4}$$

and

$$\underline{F^P} = f(\underline{F^D}) \tag{5}$$

with f defined as above and

$$\underline{F^D} \geq \Pi^{dc} - \Pi^{cc}. \tag{6}$$

If and only if

$$\underline{F^P} \leq \overline{F^P}, \tag{7}$$

then the combined punishment strategy profile $\Phi(\underline{F^D})$, where all players in both groups play strategy $\phi_{K_i}(\underline{F^D})$ defined as above, is a SPNE.

[25] With the possibility to sanction a group mate K_j whenever she chooses to defect in the first stage of Γ_{FP}, the sanctioning player K_i can get player K_j to internalize the negative externality. The sanctioning player may, thus, be able to induce player K_j to cooperate in the first stage of Γ_{FP} if the loss $F^D_{K_j}$ from the ingroup punishment is severe enough.

[26] A backward induction argument again shows that a finite repetition of Γ_{FP} under complete information would not yield cooperation in any period. In the last subgame which corresponds to the second stage of Γ_{FP}, the dominant strategy is clearly not to impose the sanction F_{K_i} since there is no gain from sanctioning to compensate the sanction costs $F^P_{K_i}$. In the next to last subgame, i.e. the first stage of Γ_{FP}, players in the outgroup pairing anticipate that there will be no ingroup punishment in the last stage. Therefore, the prisoner's dilemma (1) is played as a one-shot game with the familiar outcome (d, d). Therefore, the prediction for the last period of the two-stage game Γ_{FP} is $(d, d, 0, 0)$. This again destroys any incentives to punish in the second stage of Γ_{FP} in the period before and so on. An announcement to punish a defector is therefore not credible in a finitely repeated game under complete information (see Kreps et al., 1982, for incomplete information).

The proof is given in *Appendix A.2.* We have seen that if and only if $\underline{F^P} \leq \overline{F^P}$, the combined punishment strategy profile is a SPNE in which interethnic cooperation can be achieved by the threat of ingroup punishment. The reason for executing the ingroup punishment is to make the other group forgive the instance of defection of one's group member and thus, continue with cooperation in subsequent outgroup pairings. In order to make the other group forgive defection, the ingroup punishment must be higher than the minimum demanded by the other group, i.e. $F^D \geq \underline{F^D}$. To determine the level of $\underline{F^D}$ the other group demands at least, we reason as follows: The primary interest of each group is to get cooperation from the players of the other group. Therefore, they demand an ingroup punishment whose threat is able to induce cooperation in the first stage of Γ_{FP}. Going back to the first stage of Γ_{FP}, we can show that the ingroup punishment must satisfy $\underline{F^D} \geq \Pi^{dc} - \Pi^{cc}$.[27] If the other group observes that an ingroup punishment satisfying this condition is executed, then they forgive the defection and continue with cooperation in the subsequent outgroup pairings.

However, there is also an upper limit for the ingroup punishment to be considered which ensures that it will be applied in SPNE. A more severe ingroup punishment F also involves higher costs F^P for the punisher. The upper bound is given by (4). Since $F^P = f(F^D)$, condition (6) can be used to derive the lower bound for F^P in condition (5). If and only if parameters satisfy $\underline{F^P} \leq \overline{F^P}$, then the players from the ethnic groups will find it optimal to apply the ingroup punishment to induce interethnic cooperation.

Consider now the case where $\underline{F^P} > \overline{F^P}$. The parameters are such that there exists no SPNE where players punish their group member. In particular, this is true when $f(F^D)$ is such that the demanded minimum loss is too costly for the punisher. Imagine, for example, a situation where players value "peace" within their group much higher than "peace" with the other group. In such a situation, players will not threaten the internal peace within their group by applying ingroup sanctions in order to achieve external peace with the other group. By not punishing their defecting group member, players prefer to accept the collective outgroup punishment instead. Whether a

[27] The loss for the punished player from the ingroup punishment must be higher than the surplus of defection over cooperation in the outgroup pairing.

strategy profile with an ingroup punishment component will thus be applied by the players of the ethnic groups, primarily depends on the shape of the function $f(F^D)$.

However, even with $\underline{F^P} > \overline{F^P}$ we may get the two groups to cooperate even without ingroup punishment. This is true whenever the threat of outgroup punishment alone is strong enough to induce interethnic cooperation or, analogously to our condition in the outgroup punishment strategy profile (2), if $\Pi^{dc} - \Pi^{cc} < V^+ - V^-$. Therefore, even though the combined strategy profile is not a SPNE since the prescribed ingroup punishment is not chosen, we may still have cooperative outcomes on the equilibrium path.

Note that all $\Pi^{dc} - \Pi^{cc} \leq \underline{F^D} \leq f^{-1}(F^D) = \overline{F^P}$ are possible $\underline{F^D}$ strategy parameters that all yield interethnic cooperation with ingroup punishment. However, in terms of efficiency, the lowest possible $\underline{F^D}$ is preferred by both ethnic groups off the equilibrium path, should a group member defect ("trembling hand") at some point. We conclude that, in this sense, the best equilibrium is the one where $\underline{F^D} = \Pi^{dc} - \Pi^{cc}$ so that $\underline{F^P} = f(\Pi^{dc} - \Pi^{cc}) \leq \overline{F^P}$.

2.3 Comparison between the strategy profiles

We analysed two institutions. Each institution offers a set of possible actions for the players. The first one offers outgroup punishment by witholding cooperation in subsequent outgroup pairings. The second adds ingroup punishment in the form of a sanction F from the punisher to the punished within a group. Within each institution, a strategy profile was proposed that potentially provides interethnic cooperation on the equilibrium path.

In this section, we analyse which strategy profile is more susceptible for cooperative outcomes between ethnic groups. In particular, we ask whether adding ingroup punishment to the outgroup punishment predicts cooperative outcomes for a wider range of exogenous parameter values. Further, we specify the condition when the ingroup punishment is actually applied, when outgroup punishment alone is enough to induce cooperation or when both strategy profiles fail to induce cooperation between the two groups.

For illustration purposes, we consider strategy profiles with parameter choices of $\underline{F^D} = \Pi^{dc} - \Pi^{cc}$ only. Four different combinations of exogenous parameters may arise. We present for each regime whether the strategy profiles $\Phi(\underline{F^D})$ and Θ are SPNE.

1. $\left[V^+ - V^- \leq \Pi^{dc} - \Pi^{cc} \leq f(\Pi^{dc} - \Pi^{cc})\right] \vee \left[V^+ - V^- \leq f(\Pi^{dc} - \Pi^{cc}) \leq \Pi^{dc} - \Pi^{cc}\right]$;
   none of the strategy profiles $\Phi(\underline{F^D})$ and Θ are SPNE.

2. $f(\Pi^{dc} - \Pi^{cc}) \leq V^+ - V^- \leq \Pi^{dc} - \Pi^{cc}$;
   only $\Phi(\underline{F^D})$ is an SPNE.

3. $\left[f(\Pi^{dc} - \Pi^{cc}) \leq \Pi^{dc} - \Pi^{cc} \leq V^+ - V^-\right] \vee \left[\Pi^{dc} - \Pi^{cc} \leq f(\Pi^{dc} - \Pi^{cc}) \leq V^+ - V^-\right]$;
   both $\Phi(\underline{F^D})$ and Θ are SPNE.

4. $\Pi^{dc} - \Pi^{cc} \leq V^+ - V^- \leq f(\Pi^{dc} - \Pi^{cc})$;
   only Θ is an SPNE.

How do the two strategy profiles compare when it comes to their potential for achieving interethnic cooperation? The following lemma claims that the parameter regime with cooperative outcomes on the equilibrium path is at least as wide under strategy profile $\Phi(\underline{F^D})$ as under Θ.

Lemma 1 *If we have interethnic cooperation in equilibrium under the outgroup punishment strategy profile Θ, then we necessarily have interethnic cooperation in equilibrium under the combined punishment strategy profile $\Phi(\underline{F^D})$. However, the converse is not true.*

For the proof, we refer to *Appendix A.3.* The central force that determines whether cooperative outcomes are possible for a wider range of exogenous parameter values is the shape of the function $F^P = f(F^D)$. To be more specific, consider the linear function

$$F^P = \alpha F^D$$

with the exogenous parameter $\alpha > 0$. A low value of α indicates that executing an ingroup punishment does not involve high costs to the punisher. Clearly, we see that if we have an $\alpha > 1$, then executing an ingroup punishment causes a bigger loss to the punisher than to the punished player.

For the strategy profile $\Phi(\underline{F^D})$ to be a SPNE, we know from condition (19) that we need

$$\alpha(\Pi^{dc} - \Pi^{cc}) \leq V^+ - V^-. \tag{8}$$

For the strategy profile Θ to be a SPNE, we need

$$\Pi^{dc} - \Pi^{cc} \leq V^+ - V^-. \tag{9}$$

Comparing the conditions (8) and (9) highlights the similarities between the two punishment strategy profiles. In the outgroup punishment strategy profile, there exists no ingroup punishment. The "only" threat that the players in the outgroup pairing have to fear is the threat of a collective outgroup punishment. Through defection they will initiate an infinitely-long outgroup punishment phase and, whenever they are chosen in a subsequent outgroup pairing, they will be punished for their defection. The condition for interethnic cooperation under the outgroup punishment strategy profile (9) looks similar to the one for the combined punishment strategy profile (8). However, the nature of the threat is different. The threat that players in the outgroup pairing have to fear comes from within their respective group. The group members decide upon the ingroup punishment by comparing the long-term negative externality $V^+ - V^-$ from the defecting player of their group with the short-term cost $\underline{F^P}$ of punishing the defector within their group. The players cooperate in the first stage if the gain from defecting $\Pi^{dc} - \Pi^{cc}$ is smaller than the loss from the ingroup punishment $\underline{F^D}$. Therefore, the sanction $\underline{F^D}$ replaces the loss from an outgroup punishment phase $V^+ - V^-$ as a device to discipline a player to cooperate in an outgroup pairing. So whenever

$$\alpha(\Pi^{dc} - \Pi^{cc}) < \Pi^{dc} - \Pi^{cc}$$

or $\alpha < 1$ and

$$\alpha(\Pi^{dc} - \Pi^{cc}) \leq V^+ - V^-,$$

the set of exogenous parameters that yield cooperative outcomes is strictly larger under the combined punishment strategy profile.

Comparing the conditions for interethnic cooperation for the two strategy profiles allows us to illustrate the effect of different values for α more clearly.

Let $\widehat{\Pi} = \frac{\Pi^{dc} - \Pi^{cc}}{\Pi^{cc}}$ denote the one-time deviation gain from unilateral defection in an outgroup pairing. Recall that $\eta_o = 0.5$ under a uniform distribution sampling for the outgroup pairings with two players in each group.

Condition 2 *The condition for interethnic cooperation under the outgroup punishment strategy profile* Θ *is*

$$\delta \geq \frac{2\widehat{\Pi}}{1 + 2\widehat{\Pi}}. \tag{10}$$

That is, the discount factor must be high enough with respect to the one-time deviation gain. A higher discount factor δ could be interpreted as a shortening of the time lag between outgroup pairings. Loosely spoken, this means that players encounter players from the other ethnic group more often. This, in turn, makes an outgroup punishment strategy more effective as a disciplining device for achieving interethnic cooperation.

Condition 3 *The condition for interethnic cooperation under the combined punishment strategy profile* $\Phi(\underline{F^D})$ *is*

$$\delta \geq \begin{cases} \frac{2\alpha\widehat{\Pi}}{1+2\alpha\widehat{\Pi}} & \textit{if } 0 < \alpha < 1 \\ \frac{2\widehat{\Pi}}{1+2\widehat{\Pi}} & \textit{if } \alpha \geq 1 \end{cases}. \tag{11}$$

The condition is derived in the *Appendix A.4*. Comparing conditions (10) and (11) indicates the central role of the parameter α for evaluating the power of the two strategy profiles for achieving interethnic cooperation.

The following proposition presents the main result of our analysis.

Proposition 3 *The combined punishment strategy profile, using outgroup as well as ingroup punishment, yields more cooperation between ethnic groups than the outgroup punishment strategy profile, using outgroup punishment only. This is true in the sense that,* $\forall\alpha < 1$*, the range of exogenous parameters under which cooperative equilibria exist is strictly wider under the combined punishment strategy profile.*

Figure (1) depicts the various regimes where interethnic cooperation is possible for three different values of α. If $\alpha = 1$, then we have the border case in which players are indifferent between averting the collective outgroup punishment by executing the minimum demanded ingroup punishment and bearing the future costs of the collective outgroup punishment by not executing the necessary ingroup punishment. For higher values of α, the ingroup punishment that is demanded by the other group becomes too expensive for the punisher. She prefers to bear the collective outgroup punishment instead. Therefore, the ingroup punishment that is demanded in the combined strategy profile will not be applied. This does not mean, however, that interethnic cooperation breaks down under the combined punishment profile. The threat of the collective outgroup punishment, being a part of the combined strategy profile, is still present. In fact, for all values of $\alpha \geq 1$, the regimes of exogenous parameters that enable interethnic cooperation coincide for the outgroup and the combined strategy profile. This regime corresponds to region IV in figure (1).

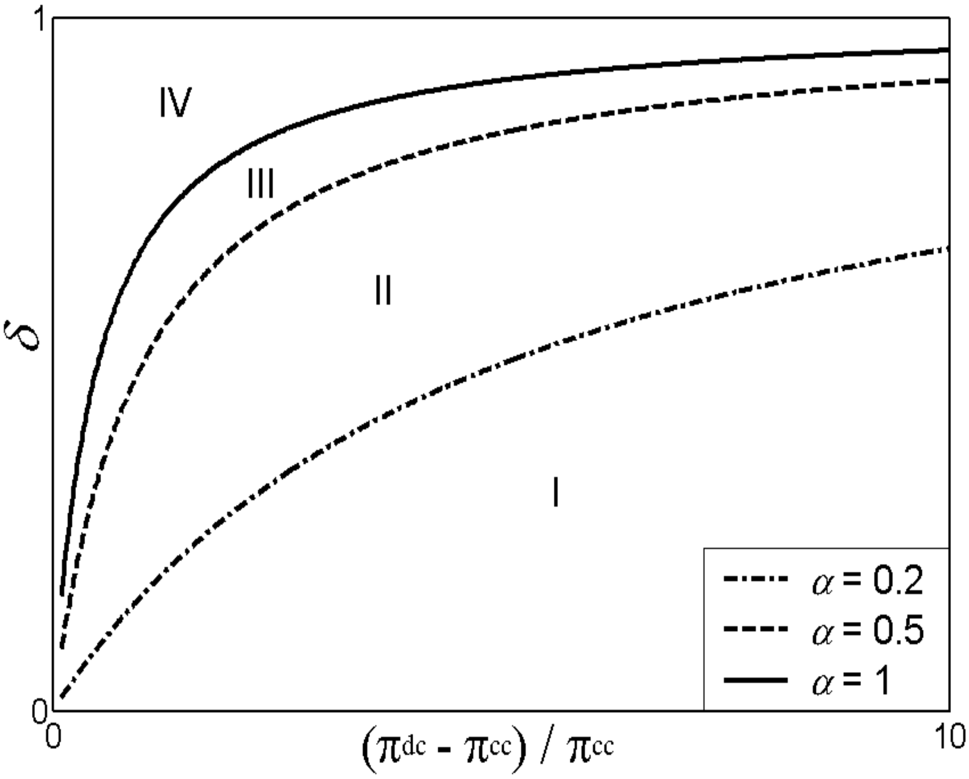

Figure 1: *The different regimes for interethnic cooperation*

The main advantage of the combined punishment strategy profile, and, in particular, the addition of ingroup punishment, becomes apparent for $\alpha < 1$. Consider, for example, region III. Here, interethnic cooperation cannot be sustained with the outgroup punishment strategy profile. However, for $\alpha = 0.5$, the costs for executing the demanded ingroup punishment are low enough so that it pays to avert the collective outgroup punishment by punishing one's defecting group member. Clearly, the set of exogenous parameters larger for the combined strategy profile (region III and IV) compared to set under the outgroup punishment strategy profile (region IV). The figure (1) also depicts the situation for $\alpha = 0.2$ with an even stronger difference between the set under the combined punishment (region II, III and IV) and the outgroup punishment (region IV) strategy profile. Region I corresponds to the set of exogenous parameters where none of the two strategy profiles is able to sustain interethnic cooperation for $\alpha = 0.2$.

We conclude this section by presenting the comparative statics of the two punishment strategy profiles in the following corollary.

Corollary 1 ***a.*** *Consider the outgroup punishment strategy profile* Θ. *Ceteris paribus,*
· *a higher discount factor* δ,
· *a lower ratio* $\frac{\Pi^{dc}-\Pi^{cc}}{\Pi^{cc}}$
raises the incentives to cooperate in outgroup pairings.
b. *Consider the combined strategy profile* $\Phi(\underline{F^D})$. *Ceteris paribus,*
· *a higher discount factor* δ,
· *a lower ratio* $\frac{\Pi^{dc}-\Pi^{cc}}{\Pi^{cc}}$
· $\forall \alpha \leq 1$, *a lower slope parameter* α *of the linear function* $F^P = \alpha F^D$
raises the incentives to cooperate in outgroup pairings.

3 Related Literature

Our paper bears relation to several strands of literature. The recent game-theoretic literature has sought to extend the result that cooperation can be achieved in infinitely repeated prisoner's dilemma games between two players if they are patient enough. The authors assumed various informational settings and analysed whether cooperation can be achieved within networks of players where the agents are randomly matched in each period. Kandori (1992) and Buskens and Weesie (1999) show that in these cases, it is not necessary that players repeatedly meet with the same partner to be able to establish a trust relationship. Whether trust works for achieving cooperative outcomes on the equilibrium path depends on the direct observability of the action choices of group members or on the way how information about the player's experiences with their partners is transmitted between the players within the network.[28] If all player's past actions are publicly observable and each player can be identified, Kandori (1992) shows that cooperative outcomes are possible for the same discount factor with an arbitrary population size and matching rule as in the two-player repeated game where the victim is able to punish the defector directly. Buskens and Weesie (1999) analyse an infinitely repeated game between one trustee and a network of trustors. In each period, either the same trustor or a new trustor is matched for the interaction with the trustee. Whether the new trustor is able to get information about the past action choices of the trustee depends on whether she has a tie with the previous trustor in the network. If there is a tie between the previous and the new trustor, then information about the past action choices of the trustee is communicated and the new trustor is able to condition her play on this information. Depending on how dense the network of the trustors is, i.e. how many ties there are between the players, the better a decentralized punishment institution works to induce the trustee to cooperate, i.e. to honor trust. Even if no information can be transmitted and players are not able to identify other players, there exist equilibria with cooperative outcomes such as contagious equilibria where players defect whenever they experienced an incidence of defection in one of their past matches (Kandori, 1992, Ellison,

[28] A rather modern form of financing small-scale projects used by the Grameen Bank in Bangladesh uses the advantage that players within networks know each other's action choices better than an outside party like a bank. They are more efficient in punishing their group members for failing to contribute to the group's interest. In this form of financing, the bank holds the entire group jointly liable for repaying the debt of any member of their group and therefore, it is in the public interest of the group that every member pays back her debt. Ingroup punishment is applied if a player is able to repay her part but decides to free ride instead (Besley, Coate, 1995, La Ferrara, 1999).

1994).

All these models are highly valuable for analysing interactions within groups. In particular, the network advantages inherent in ethnic groups which come close to a situation of perfect public observability pose an ideal setting for applying the community enforcement mechanism proposed, for example, by Kandori (1992). However, the models have to be adapted to analyse interactions across groups which are characterised by the special informational setting we impose in our model where we have full identification within the ethnic group[29] but no identification of players across groups.

An interesting approach is offered by Bowles and Gintis (2004) who use an evolutionary model to explain the boundaries of ethnic groups. In their model, the foregone opportunities by not trading with outgroup players are offset by an enhanced ability to solve problems of incomplete contracts through a higher level of trust among ingroup players. The analysis of this trade-off allows them to explain the evolution of persistent parochialism of ethnic groups.[30]

Our model is perhaps closest to Fearon and Laitin (1996). They analyse an institution with in-group policing which uses ingroup punishment alone and a spiral regime which uses outgroup punishment alone. However, in their model, the execution of ingroup punishment is assumed to be costless for the punisher. We believe that the introduction of an ingroup punishment which is not only costly to the punished but to the punishing player as well is more realistic. In their model, the introduction of such an ingroup punishment would lead to a breakdown of their in-group policing equilibrium because the incentives would be such that no ingroup punishment would ever be applied by rational agents.

Further, we focus on prisoner's dilemma games between ethnic groups only whereas Fearon and Laitin (1996) model intra- and interethnic prisoner's dilemmas simultaneously. However, this simplification does not weaken our

[29] The bigger the groups, the less plausible the full information assumption within groups appears. So in this respect, our model of decentralised ingroup punishment works best in small groups.

[30] See also McElreath et al. (2003) for the evolution of ethnic markers as a response to a coordination game.

arguments since our focus is on interethnic interactions and we subsumed all possible effects of intra- on interethnic interactions in our reduced-form function of ingroup punishment. The benefits of this simplification are that it allows us to analyse both ingroup and outgroup punishment simultaneously in a tractable manner. In the institution with outgroup punishment alone, the only punishment that a player in the outgroup pairing may expect comes from the other group. In the combined punishment strategy, the punishment either comes from within the group or from the other group. The analysis of the second institution can be seen as the major novel contribution of this model. To our knowlegde, no analysis existed so far that provided such a simultaneous treatment. The simultaneous inclusion of ingroup and outgroup punishment yields new insights in the mechanisms that characterise a large number of interethnic interactions as we believe that most interethnic interactions are embedded in an institution with elements of ingroup and outgroup punishment at the same time. Fearon and Laitin (1996) suspect a natural tendency in their spiral regime (which corresponds to our outgroup punishment strategy profile) towards the development of features of ingroup punishment but they did not pursue this idea more formally. In our model, we provide such a theoretical analysis and show that the inclusion of ingroup punishment in addition to outgroup punishment alone yields cooperation on the equilibrium path under a wider exogenous parameter regime.

An interesting contribution is provided by Greif who uses game-theoretic models to explain the development of institutions in history. In Greif (1997), he provides a historical analysis on the use of collective punishment in medieval cities where a community was held responsible for a defection of one of their citizens. In another work, Greif et al. (1994) emphasize the role of community enforcement through guilds in the development of long-distance trade in the 12th century which underlines the relevance of ingroup punishment.[31] Examples of ingroup punishment are also found in the Ottoman Empire where a millet system was in place. Each group could live in autonomy as long as their millet leader was successful in monitoring and policing her group members in interactions with outsiders (Fearon and Laitin, 1996,

[31] See also Bardhan (1999) for a good review.

Dumont, 1982). Comparable to the millet system in the Ottoman Empire is the organization of the Hausa community in Yorubaland of Nigeria (Cohen, 1969). The Hausa community used a very strong form of ingroup punishment executed by the Hausa Chief which served as a disciplining device to induce cooperation not only among the group members but also towards outsiders. Interestingly, this could be the reason why the Hausa traders were so successful in maintaining the trade monopoly in Yorubaland.

These works provide an empirical backing of our analysis that the forms of outgroup and ingroup punishment we use in our model have actually been applied in history. In particular, Greif's contributions underline the importance of punishment institutions for the development of markets in general. However, not much research has been provided so far or as Greif (1993, p.525) puts it in his introduction: "Yet not so much is known about the historical institutional developments that enabled exchange relations to expand, even though such knowledge can shed light on the nature and evolution of modern institutions and facilitate the understanding of the institutional transitions that developing economies still face."

4 Conclusion

In this paper, we model interethnic conflicts as a possible response to an informational problem which is inherent in interactions between players of different ethnic groups. We assume that players are able to observe the action choices of each outgroup pairing. However, whereas the players know which particular player in their own group was responsible for the action choice, they do not know who was responsible for the action choices in the other group. This means that we have an identification problem across groups. As a consequence, players cannot regress on punishment strategies which are targetted on potential defectors in the other group to induce cooperation in outgroup pairings.

This assumption may appear unrealistic given our model with only two players in each group. It is indeed a simplification that allows us to ignore the free rider problem of the second-order public good of ingroup punishment (Boyd and Richerson 1992, Panchanathan and Boyd 2004). In a model with

$n > 2$ players in each group we would have to tackle the delicate question of who punishes the potential defector in the same group. The execution of ingroup punishment can be seen as a second-order public good. Each group member has an incentive to free ride by letting the others carry the costs of executing the ingroup punishment themselves.[32] However, in this model, we want to concentrate on the potential power of ingroup punishment strategies for inducing cooperation with outgroup players. One might imagine, however, that in a group with $n > 2$, the costs for ingroup punishing a potential defector may be spread among the other ingroup members.[33] In fact, this would lower the individual costs for executing the minimal necessary ingroup punishment and we should observe cooperation for an even wider range of exogenous parameters. So, in that sense, our analysis provides a lower bound for achieving cooperation across groups.

We propose two strategy profiles that are able to induce interethnic cooperation in a SPNE. In the first strategy profile, we consider a punishment that is only applied by players of the other group. Here, each player indiscriminately punishes her partner in all subsequent outgroup pairings for a defection from an arbitrary player from the partner's group. Since it is a best response to this punishment strategy of the victim's group member to defect as well, we have a situation where one incidence of a defection is enough to trigger an interethnic conflict where all players of both groups defect against each other for infinitely many periods. More figuratively, this can be interpreted as an interethnic conflict within our model.

We have seen that such an incidence of defection of an arbitrary player of an arbitrary group creates a huge negative externality, not only on the players of the other group but also on her own group members. Therefore, a natural

[32] The public goods problem explains why in very large groups like nations, formal state authorities take over the role of ingroup punishment as opposed to a decentralised system in small groups. The taxes imposed on the group members to finance the ingroup punishment can thus be interpreted as the punisher's costs of ingroup punishment. More realistically, we are embedded in a system of ingroup punishment with formal as well as informal elements.

[33] However, problems of diffused responsibility may arise in that every group member expects the others to execute the ingroup punishment on the defector. The famous case of the murder of Kitty Genovese in 1964 (see, for example, Dixit and Skeath, 1999, on pp.389), is one tragic example of this phenomenon.

response would be that ingroup members might consider to discipline their ingroup members by some form of ingroup punishment as well. We model this ingroup punishment as a choice F. That is, we use a reduced form approach. In a second strategy profile, we set up a new game which adds the option of ingroup punishment to the original game as a second stage decision. A potential defector now gets punished either by her own group member or by the players from the other group whenever she is matched in subsequent outgroup pairings. Which players execute the punishment depends on the relative costs of executing the punishment. The punisher's costs of ingroup punishment are a monotonically increasing function of the loss that the punisher wants to induce to the potential defector in her group. The advantage of ingroup punishment over the collective outgroup punishment is that it can be targetted. Whether ingroup punishment is executed (it is only executed if it is better from an individual point of view; compare to whether it should be executed from a group's point of view) depends on whether the future value of cooperation (Π^{cc}) over defection (Π^{dd}) in outgroup interactions is large enough, whether etc.

We have shown that the second strategy profile yields cooperation for a wider range of exogenous parameters depending on the function $f(F)$. More specifically, this is true for all ingroup punishments where the ingroup punishment involves less costs for the punisher than for the punished player.

However, one has to be careful with praising the ingroup punishment device for its ability to induce interethnic cooperation. Ingroup punishment is only a disciplining device that serves players to enforce norms within their groups. It is an efficient way of enforcing such a norm in the sense that the decision about the punishment lies in the hands of individuals which are able to target the punishment on particular players.

We implicitly assumed that there is a preference for mutual cooperation in outgroup pairings given by the assumption that $\Pi^{dd} < \Pi^{cc}$. However, one might imagine situations where the two groups are at war entailing so that players not only prefer the outcome Π^{dc} over Π^{cc} but they also value $\Pi^{dd} > \Pi^{cc}$. The ingroup punishment could then be used in such a case to induce group members to defect against players from the other group, i.e. to enforce a defection norm.[34] Therefore, the effect of the institution of ingroup punishment on the level of cooperation depends on the attitude of the groups towards each other. Whether the cooperation norm or the defection norm is in place depends on the particular situation.[35]

The question on how group attitudes are formed is itself an interesting research question which we did not pursue in this model. A clear understanding of the dynamics involved demands a closer look on the political process and whether the ingroup punishment is decentralised (e.g. neighbour control), centralised (by some form of state authority) or both. One could think, for example, that those with a stronger interest in enforcing their desired group norm have lower costs in enforcing the ingroup punishment. The cooperation decision of an individual player then depends on the extent of ingroup punishment she expects following a defection or a cooperation decision.

The model helps politicians who seek to maintain peace between the ethnic groups of their state. They can put additional pressure on the leaders of the ethnic groups who are designated to be responsible for the execution of ingroup punishment within their group. So, in some ways, they can form the norm of the ethnic group through sanctioning devices on the leaders of the ethnic groups.[36]

[34] Bornstein (2003) reviews the literature on intergroup conflicts where the groups interact in a team game. Each player has to decide how much to contribute to the group's effort. In the situation of step-level public goods, the group with the higher number of contributors wins the game. In the situation with a continuous public good, the amount by which the number of contributors surpasses the number of contributors of the other group is important as well. Interestingly, there is interest conflict between the group's interest and the collective interest of both groups. The dominant strategy for an individual is not to contribute. However, for the group's interest, the more contributors there are in a group, the better it is. Finally, the less contributors there are on both sides, the better the situation for both groups.

[35] In our model, group members that do not follow the outgroup punishment strategy are not punished by their group members. However, through fear of retaliation by her own group members, it may be the case that players choose to punish players from the other group. One could even think that during an outgroup punishment phase, the cooperation norm essentially becomes a defection norm and each player that does not participate in defecting against her partner in the outgroup pairings gets punished by her ingroup members. Despite the different motive behind defecting during an outgroup punishment phase, the outcome would be the same as with the strategy profile we propose in our model.

[36] This institutional solution was applied in the Ottoman Empire for instance (Dumont,

Our model has been framed in the context of interethnic relations. In our opinion, this simplified the presentation of the main results by making them more intuitive as well as more figurative for the reader. However, the areas where our model is applicable are not restricted to interethnic interactions only. One can, for instance, translate the findings to shed light on how to achieve cooperation between departments within a firm or between international trade partners. More generally, our model offers insights for all settings where interactions between groups are involved whose group members are connected by some form of collective reputation.[37]

References

Bardhan, P. (1999): "Distributive Conflicts, Collective Action, and Institutional Economics", *University of California, Berkeley.*

Besley, T., Coate, S. (1995): "Group Lending, Repayment Incentives and Social Collateral", *Journal of Development Economics*, 46, 1-18.

Boehm, C. (1994): "Blood Revenge: The Anthropology of Feuding in Montenegro and Other Tribal Societies", *Lawrence: University Press of Kansas.*

Bowles, S., Gintis, H. (2004): "Persistent parochialism: trust and exclusion in ethnic networks", *Journal of Economic Behavior and Organisation*, 55, 1-23.

Boyd, R., Richerson, P.J. (1992): "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups", *Ethology and Sociobiology*, 13, 171-195.

Bornstein, G. (2003): "Intergroup Conflict: Individual, Group and Collective Interests", *Personality and Social Psychology Review*, 7, 129-145.

Brewer, M. (1979): "In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis", *Psychological Bulletin*, 86, 307-324.

Buskens, V., Weesie, J. (1999): "Cooperation via Social Networks", *Utrecht*

1982).

[37] See Tirole (1996) on a model of collective reputation with incomplete information.

University (mimeo).

Cohen, A. (1969): "Custom and Politics in Urban Africa", *Berkeley: University of California Press.*

Colson, E. (1974): "Tradition and Contract", *Chicago: Aldine.*

Dixit, A., Skeath, S. (1999): "Games of Strategy", *W.W. Norton & Company.*

Dumont, P. (1982): "Jewish Communities in Turkey during the Last Decades of the 19th Century", in: Christians and Jews in the Ottoman Empire, *vol.1*, Braude, B., Lewis, B. (eds.), *New York: Holmes and Meier.*

Ellison, G. (1994): "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching", *Review of Economic Studies*, 61, 567-588.

Falk, A., Fehr, E., Fischbacher, U. (2001): "Driving Forces of Informal Sanctions", *Working paper of the Institute for Empirical Research in Economics, University of Zurich*, No. 59.

Fearon, J.D., Laitin, D.D. (1996): "Explaining Interethnic Cooperation", *American Political Science Review*, 90, 715-735.

Fehr, E., Fischbacher, U. (2004): "Third-party punishment and social norms", *Evolution and Human Behavior*, 25, 63-87.

Green, E., Porter, R. (1984): "Noncooperative Collusion under Imperfect Information", *Econometrica*, 52, 87-100.

Greif, A. (1993): "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition", *The American Economic Review*, 83, 525-548.

Greif, A., Milgrom, P., Weingast, B. (1994): "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild", *Journal of Political Economy*, 102, 745-776.

Greif, A. (1997): "On the Social Foundations and Historical Development of Institutions that Facilitate Impersonal Exchange: From the Community Responsibility System to Individual Legal Responsibility in Pre-modern Europe"", *Stanford Working Paper,* No. 97-016.

Halberstadt, A., Lik Mui, M.M. (2002): "A Computational Model of Trust and Reputation", *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Hasluck, M. (1954): “The Unwritten Law in Albania”, *Cambridge: Cambridge University Press.*

Horowitz, D. (1998): “Structure and Strategy in Ethnic Conflict”, *Paper prepared for Annual World Bank Conference on Development Economics.*

Kandori, M. (1992): “Social Norms and Community Enforcement”, *Review of Economic Studies*, 59, 63-80.

Kreps, D., Milgrom, P., Roberts, J., Wilson, R. (1982): “Rational Cooperation in the Finitely Repeated Prisoner’s Dilemma”, *Journal of Economic Theory*, 27, 245-252.

La Ferrara, E. (1999): “Ethnicity and Reciprocity: A Model of Credit Transactions in Ghana”, *University of Bocconi* (mimeo).

McElreath, R., Boyd, R., Richerson P.J. (2003): “Shared Norms and the Evolution of Ethnic Markers”, *Current Anthropology*, 44, 1, 122-129.

Unesco (1974): “Two studies on ethnic group relations in Africa: Senegal and the United Republic of Tanzania”, *Unesco.*

Panchanathan, K., Boyd, R. (2004): “Indirect reciprocity can stabilize cooperation without the second-order free rider problem”, *Nature*, 432.

Raub, W., Weesie, J. (1990): “Reputation and Efficiency in Social Interactions: An Example of Network Effects”, *American Journal of Sociology*, 96, 626-654.

Rotemberg, J., Saloner, G. (1986): “A Supergame-Theoretic Model of Business Cycles and Price Wars During Booms”, *American Economic Review*, 76, 390-407.

Tirole, J. (1996): “A Theory of Collective Reputations (with applications to the persistence of corruption and to firm quality)”, *Review of Economic Studies*, 63, 1-22.

Yamagishi, T., Kiyonari, T. (2000): “The Group as the Container of Generalized Reciprocity”, *Social Psychology Quarterly*, 63, 116-132.

A. Appendix

A.1 Proof of Proposition 1

For simplicity, we set $\Pi^{dd} = 0$. Let

$$V^{+} = \textstyle\sum_{t=1}^{\infty} \delta^{t} \eta_{o} \Pi^{cc}$$

or, by convergence of the geometric sequence,

$$V^{+} = \frac{\delta}{1-\delta} \eta_{o} \Pi^{cc} \tag{12}$$

denote the discounted payoff of an outgroup cooperation phase, where all players in the outgroup pairings cooperate, starting in the subsequent period. Further, because of $\Pi^{dd} = 0$, we have

$$V^{-} = 0 \tag{13}$$

for the discounted payoff of an outgroup defection phase, where all players apply an OGP whenever matched in an outgroup pairing.

First, we consider all subgames of class III. Given that the other player plays strategy θ in the outgroup pairing, the best response is to defect as well since $\Pi^{cd} < \Pi^{dd}$ by assumption.

Second, we consider subgames of classes I and $II.$ and $III.$ To induce an arbitrary player to cooperate in an outgroup pairing, the short-term gain from defecting must be lower than the long-term loss from an outgroup punishment phase. Hence, the "no deviation" constraint must satisfy

$$\Pi^{dc} - \Pi^{cc} \leq V^{+} - V^{-} \tag{14}$$

or, rearranging and substituting for $V^{+}$and V^{-},

$$\frac{\Pi^{dc} - \Pi^{cc}}{\Pi^{cc}} \leq \frac{\delta}{1-\delta} \eta_{o}$$

which concludes the proof.

A.2 Proof of Proposition 2

The only purpose of executing an IGP is to make the players of the other group forgive the defection of one's group member. All IGP choices of $F \geq \underline{F}$ satisfy this forgiveness requirement. Since higher choices of F are more costly to the punishing player, it makes no sense to choose any $F > \underline{F}$. If a player decides to punish at all, she will only consider an IGP of $F = \underline{F}$.

Ingroup interactions. First, consider subgames of class $\widetilde{I}$. In such a case, the best response is indeed not to apply any IGP since there are no gains but only costs of F^P to the punishing player from choosing $F > 0$. More particularly, $-\underline{F^P} + V^+ < V^+$. Second, consider subgames of class $\widetilde{II}$. An IGP of $F = \underline{F}$ causes a loss of $\underline{F^P}$ to the punishing player. It makes the players from the other group forgive the defection of one's group member. The loss from executing the IGP has to be smaller than the gain $V^+ - V^-$. The ingroup member of the defector chooses to punish her group member in the second stage of Γ_{FP} if

$$-\underline{F^P} + V^+ \geq V^-$$

or, after rearranging and substituting for V^- and V^+,

$$\underline{F^P} \leq \frac{\delta}{1-\delta}\eta_o \Pi^{cc} = \overline{F^P}.$$

Finally, consider subgames of class $\widetilde{III}$. Given strategy $\phi_{K_i}(\underline{F^D})$, there is no point in apply an IGP during an outgroup punishment phase since it does not lead to any kind of forgiveness or any change in subsequent action choices by the players of the other group. More particularly, the payoff from executing the minimum IGP during an outgroup punishment phase $-\underline{F^P} + V^-$ thus will always be greater than the payoff V^- without any IGP.

Outgroup interactions. Consider subgames of classes I or II. There are two situations that need to be considered. First, consider a situation where the parameters are such that a player can anticipate that no IGP will be executed on her if she chooses to defect.[38] She cooperates if and only if

$$\Pi^{dc} - \Pi^{cc} \leq V^{+} - V^{-} \tag{15}$$

or, substituting for $V^{+}$and V^{-},

$$\Pi^{dc} - \Pi^{cc} \leq \frac{\delta}{1-\delta}\eta_o \Pi^{cc}$$

which is analogous to the condition for subgame perfection of strategy profile Θ with OGP alone in game Γ_F. Next, consider a situation where a player can anticipate that she will be punished by her group member when she defects.[39] The condition for cooperation with the threat of a subsequent IGP is

$$\Pi^{cc} + V^{+} \geq \Pi^{dc} - \underline{F^D} + V^{+}$$

or

$$\underline{F^D} \geq \Pi^{dc} - \Pi^{cc}. \tag{16}$$

When players from a group forgive a defection as a result of an IGP, they abstain from an OGP in all subsequent outgroup pairings. The only purpose of forgiving lies in the incentives it creates for the group member of a defector and for the player in the outgroup pairing. First, it induce the group member to execute an IGP on defector and to carry the cost of punishment of defectors. Second, the anticipation of such an IGP creates the necessary incentives for cooperation in the first stage of the game. In order to give the proper incentives for the player in the outgroup pairing, the minimal IGP that is demanded by the players from the other group has to be strong enough to induce cooperation in the first stage of the game. In this sense, constraint (16) gives the lower bound of the strategy parameter $\underline{F^D}$ that still accomplishes this goal. This lower bound can then be translated into

$$\underline{F^P} = f(\underline{F^D})$$

[38] This is true when the costs of IGP for the punishing player are very high (?).

[39] More particularly, this is the case when $F^D < F^P$.

which is the lower bound for the loss that has to be incurred by the punishing player in order to make the players from the other group forgive. However, we have seen in the first part of the proof that $\underline{F^P}$ is also bounded from above by

$$\underline{F^P} = f(\underline{F^D}) \leq \frac{\delta}{1-\delta}\eta_o \Pi^{cc} = \overline{F^P}$$

to make an IGP choice of $F = \underline{F}$ a best response given our strategy $\phi_{K_i}(\underline{F^D})$. Therefore, for the strategy profile $\Phi(\underline{F^D})$ to be a SPNE, we must have

$$\underline{F^P} \leq \overline{F^P}.$$

Finally, consider subgames of class III. Given the other player's strategy $\phi_{K_i}(\underline{F^D})$, it is a best response to defect as well so that mutual defection is indeed the Nash equilibrium prediction in all these subgames.

A.3 Proof of Lemma 1

Consider the first part. In case 1, none of the strategy profiles is able to achieve interethnic cooperation. In case 3, both strategy profiles achieve interethnic cooperation on the equilibrium path. Strategy profile Θ uses the threat of outgroup punishment whereas strategy profile $\Phi(\underline{F^D})$ relies on ingroup punishment. Case 4 is special. Here, only strategy profile Θ is an SPNE. The costs of punishing a group member are too high relative to the long-term gain from avoiding an outgroup punishment ($\underline{F^P} > \overline{F^P}$).[40] Therefore, it is not optimal to execute an ingroup punishment in the subgames of class $\widetilde{II}$ and, as indicated above, the strategy profile $\Phi(\underline{F^D})$ is not an SPNE. However, this does not mean that strategy profile $\Phi(\underline{F^D})$ is not able to induce cooperation. We still have interethnic cooperation under strategy profile $\Phi(\underline{F^D})$ in case 4 since the threat of the outgroup punishment is strong enough

$$\Pi^{dc} - \Pi^{cc} \leq V^{+} - V^{-}$$

[40] This could have been the constellation in Palestinian conflict where arresting the terrorists on the Palestinian side was too expensive - either politically or technically - for PLO leader Arafat.

to do the job of inducing cooperation in the first stage of Γ_{FP}.

Consider the second part. Case 2 is the case where the converse is not true. Only strategy profile $\Phi(\underline{F^D})$ is able to achieve interethnic cooperation.

A.4 Proof of condition 3

The condition for cooperation with the combined punishment strategy is derived as follows. We know that the combined strategy profile $\Phi(\underline{F^D})$ is a SPNE if and only if $\underline{F^P} \leq \overline{F^P}$. Then there exists an equilibrium with cooperative outcomes and IGP in the case of a defection. In order to check this inequality, two conditions have to be met. The first one is the condition for cooperation

$$\underline{F^D} \geq \Pi^{dc} - \Pi^{cc}$$

given that an IGP will be executed. Using the linear function $\underline{F^P} = \alpha \underline{F^D}$, we get the lower bound for the costs that the punisher has to bear to induce the player in the outgroup pairing to cooperate in the first stage. Suppose, for simplicity, that $\underline{F^D}$ is chosen to be equal to the lower bound $\Pi^{dc} - \Pi^{cc}$.[41] This allows to write

$$\underline{F^P} = \alpha(\Pi^{dc} - \Pi^{cc}). \tag{17}$$

Next, we know that there is also an upper bound for the costs for the punisher

$$\underline{F^P} \leq \frac{\delta}{1-\delta}\eta_o \Pi^{cc} = \overline{F^P} \tag{18}$$

which is the second condition. That is, $\underline{F^P}$ can be as high as $\overline{F^P}$ and still we have a cooperative equilibrium with the threat of an ingroup punishment. Higher values of $\underline{F^P}$ lead to a breakdown of the equilibrium that uses an ingroup punishment as a disciplinary mechanism. Combining conditions (17) and (18), we get

$$\alpha(\Pi^{dc} - \Pi^{cc}) \leq \frac{\delta}{1-\delta}\eta_o \Pi^{cc}. \tag{19}$$

[41] This corresponds to the most preferred strategy parameter by both groups since it involves the lowest possible IGP costs for the punishing players that still induce interethnic cooperation.

Recall that $\eta_o = \frac{1}{2}$. Rearranging (19) gives condition (11) for cooperation with ingroup punishment

$$\delta \geq \frac{2\alpha\widehat{\Pi}}{1 + 2\alpha\widehat{\Pi}}$$

where $\widehat{\Pi} := \frac{\Pi^{dc} - \Pi^{cc}}{\Pi^{cc}}$.

However, interethnic cooperation is possible under strategy profile $\Phi(\underline{F^D})$ even if it is not a SPNE. Consider the case where the exogenous parameters are such that

$$\frac{\delta}{1-\delta}\eta_o\Pi^{cc} < \alpha(\Pi^{dc} - \Pi^{cc}).$$

This means that it is not a best response to execute an IGP in the subgames of class $\widetilde{II}$ and the strategy profile $\Phi(\underline{F^D})$ is not a SPNE. Still, interethnic cooperation is possible on the equilibrium path as long as the threat of the OGP is strong enough. More specifically, we need

$$\Pi^{dc} - \Pi^{cc} \leq \frac{\delta}{1-\delta}\eta_o\Pi^{cc}$$

or, by rearranging,

$$\delta \geq \frac{2\widehat{\Pi}}{1 + 2\widehat{\Pi}}. \tag{20}$$

This allows to conclude that, whenever $\alpha \geq 1$, the condition for interethnic cooperation under the strategy profile $\Phi(\underline{F^D})$ becomes (20).