

# Dynamic Group Formation in the Repeated Prisoner's Dilemma<sup>1</sup>

Toshimasa Maruta<sup>2</sup> and Akira Okada<sup>3</sup>

February, 2010

ABSTRACT: We consider dynamic group formation in repeated  $n$ -person prisoner's dilemma. Agreements in coalitional bargaining are self-binding in that they are supported as subgame perfect equilibria of repeated games. Individuals are allowed to renegotiate the cooperating group agreement through a process of voluntary participation. We prove that a cooperating group forms as an absorbing state of a Markov perfect equilibrium after a finite number of renegotiations if and only if the group is efficient, provided that individuals are patient. The cooperating group can only expand.

*Journal of Economic Literature* Classification Numbers: C70, C72, D70.

KEYWORDS: group formation, prisoner's dilemma, repeated games, non-cooperative coalitional bargaining, efficiency, renegotiation

---

<sup>1</sup>We are grateful for useful comments from seminar participants on previous versions of the manuscript. Maruta gratefully acknowledges financial support from JSPS Grant-in-Aid for Scientific Research (C)20530161. Okada gratefully acknowledges financial support from JSPS Grant-in-Aid for Scientific Research (S)20223001.

<sup>2</sup>Advanced Research Institute for the Sciences and Humanities, Nihon University, 12-5 Goban-cho, Chiyoda, Tokyo 102-8251, Japan. E-mail: [maruta.toshimasa@nihon-u.ac.jp](mailto:maruta.toshimasa@nihon-u.ac.jp)

<sup>3</sup>Corresponding author: Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601 Japan. E-mail: [aokada@econ.hit-u.ac.jp](mailto:aokada@econ.hit-u.ac.jp)

# 1 Introduction

Group formation is a complex and dynamic process in social, economic and political situations. It involves various types of negotiations among many players. In most cases, a game of group formation is repeated rather than one-shot: the players form and reform groups through a dynamic process of repeated negotiations.

The history of the EU provides one example. In 1951, six countries (Belgium, France, Germany, Italy, Luxembourg and the Netherlands) initiated the European Coal and Steel Community. Several expansions have taken place since then, and the EU currently comprises 27 member countries. Another good example is the ongoing international negotiation over climate change. In 1997, nearly all developed countries signed the Kyoto Protocol, committing to reduce emissions by 5.2 per cent (below 1990 levels) between 2008 and 2012. A renegotiation of the Protocol is now (in 2009) taking place, one of the main issues being whether or not all other countries should join the agreement.

In this paper, we consider dynamic group formation in a repeated,  $n$ -person prisoner's dilemma.<sup>1</sup> In addition to its empirical interest, group formation is an important theoretical issue in repeated games. In the theory of repeated games, the folk theorem states that if individuals are patient, every group of cooperators whose members are all better off than they would be in the defection equilibrium can be sustained as a subgame perfect equilibrium of the repeated prisoner's dilemma (Fudenberg and Maskin 1989). A well-known drawback of the folk theorem is that the set of equilibrium outcomes is plethoric. The largest group may be possible in equilibrium, but many smaller groups are also possible. In a partial-cooperation equilibrium, some players form a group to cooperate while other players free-ride on the group.

It is not yet clear how the conflict between group members and free riders can be resolved in the model of repeated games. Any suitable equilibrium selection theory should solve this group formation problem. A common practice in applied works is to focus analysis on the full-cooperation equilibrium, or else that with the largest possible

---

<sup>1</sup>Following standard terminology in the theory of collective actions, we use 'group' instead of 'coalition' to describe an association of agents.

group of cooperators, by bringing in additional conditions such as efficiency, symmetry, and/or a focal point. However, it is not clear that the largest group of cooperators actually forms when every individual has an incentive to defect. Another limitation of focusing on the full-cooperation equilibrium is the possibility of committed agents. For example, suppose that one player commits himself not to cooperate prior to game play, and that given his decision all other individuals find it beneficial to cooperate. Thus, each player expects all others to cooperate should he defect, and each member has an incentive to deviate from the largest group. In this sense, the largest group may not be stable.

This paper considers the problem of group formation in the repeated prisoner's dilemma. At the beginning of each period, players negotiate a self-binding agreement defining the strategy of the cooperating group. The agreement is assumed to be renegotiable in every period. The threat point of renegotiation is the current agreement. A new group, if any, must include the current one. Since the prisoner's dilemma describes an anarchic situation, where no individuals are forced to join any collective action, voluntary participation is a critical factor (Dixit and Olson 2000). Our coalitional bargaining model starts with voluntary participation. We will prove that a cooperating group forms as an absorbing state of a Markov perfect equilibrium in a finite number of renegotiations if and only if it is efficient, provided that individuals are patient. The theorem implies that an efficient group of cooperators necessarily forms through successive negotiations.

This paper is closely related to recent works on dynamic coalition formation (Seidmann and Winter 1998, Okada 2000, Gomes 2005, Gomes and Jehiel 2005, Bloch and Gomes 2006, Hyndman and Ray 2007 among others). On the subject of negotiations on coalition formation without externality, described as super-additive characteristic function games, Seidmann and Winter (1998) and Okada (2000) show that renegotiations necessarily result in the formation of the grand coalition in a Markov perfect equilibrium under the conditions that players are patient and coalitions can only expand. Their efficiency theorem has been extended in various directions. Gomes (2005) produces the same result for partition function games with externality. Gomes and Jehiel (2005) develop

a general set-up where coalitions may break up, and identify a necessary and sufficient condition that guarantees the convergence to efficiency. The condition is the existence of an efficient state that is free of negative externality. Hyndman and Ray (2007) extend the efficiency result to non-Markov perfect equilibria for characteristic function games. Bloch and Gomes (2006) examine the role of outside options and establish the efficiency result when there exists no externality on outside options. For an approach different from ours, Konishi and Ray (2003) consider coalition formation as a Markov process from the viewpoint of cooperative game theory.

A key difference between this paper and the prior literature is that we combine the theory of dynamic coalition formation with that of repeated games. This approach adds several new ingredients. First, it enables us to apply the theory of dynamic coalition formation to a wide range of economic situations described as strategic-form games, including the prisoner's dilemma (the subject of this paper). Previous works mainly deal with coalitional-form games with and without externality, which are not always suitable for describing strategic behavior. For example, a key notion of an efficient negative externality-free state in Gomes and Jehiel (2005) does not exist in the prisoner's dilemma, where every individual is free to defect.<sup>2</sup> Second, most previous studies assume that agreements are binding (at least temporarily) in the sense that once agreed, they are enforced by an outside mechanism, subject to renegotiations in future periods. We, on the other hand, consider *self-binding* agreements that are supported as subgame perfect equilibria of repeated games. In our approach, individuals negotiate (and renegotiate) for self-binding repeated-game profiles such as trigger strategies.

The paper is organized as follows. Section 2 defines the notation and reviews some preliminary results. Section 3 presents a model of dynamic group formation in the repeated prisoner's dilemma. Section 4 presents our main theorem. Section 5 concludes the paper.

---

<sup>2</sup>Suppose that all individuals cooperate. If any individual defects, then all other cooperators become worse off. That is, the individual's deviation causes negative externality on other players.

## 2 Preliminaries

Consider the following  $n$ -person prisoner's dilemma game. Let  $N = \{1, 2, \dots, n\}$  be the set of players. Every player  $i \in N$  independently chooses one of two actions,  $C$  (cooperation) or  $D$  (defection). All players are symmetric, with the payoff function

$$u(a_i, h), \quad a_i = C, D, \quad h = 0, 1, 2, \dots \quad (2.1)$$

where  $a_i$  is player  $i$ 's own action and  $h$  is the number of other players who select  $C$ .<sup>3</sup> The payoff function  $u$  satisfies

**Assumption 2.1.** (1)  $u(D, h) > u(C, h)$  for every  $h$ , (2)  $u(C, n - 1) > u(D, 0)$ , (3)  $u(C, h)$  and  $u(D, h)$  are increasing in  $h$ .

This assumption is standard in the literature (Schelling 1978). Property (1) means that defection is the dominant action for every player; that is, every individual is better off defecting than cooperating, regardless of the others' actions. Thus, the action profile  $(D, \dots, D)$  is a unique Nash equilibrium of the game. This equilibrium will be referred to as the defection equilibrium. On the other hand, property (2) means that if all  $n$  players cooperate, each of them is better off than they would be in the defection equilibrium. Property (3) means that cooperative action has positive externality on all players. Let  $G$  denote this game.

For  $S \subseteq N$ , let  $a^S = (a_i^S)_{i \in N}$  denote the action profile in which  $a_i = C$  for every  $i \in S$  and  $a_i = D$  for every  $i \notin S$ . For  $S \subseteq N$ , let  $s$  denote the size of  $S$  whenever no confusion arises. The same convention applies to alternative subsets of the same type, such as  $T$ ,  $a^T$ , and  $t$  below. We now introduce some key notions in our analysis.

**Definition 2.1.** Let  $S \subset T \subseteq N$ .

(1)  $T$  is called a *cooperative group given  $S$*  if

$$u(C, t - 1) > u(D, s). \quad (2.2)$$

---

<sup>3</sup>For convenience of analysis, the payoff function  $u$  is defined for all non-negative integers  $h$ .

When  $S = \emptyset$ ,  $T$  is simply called a *cooperative group*. A cooperative group  $S$  is called a *maximal cooperative group* if there exists no cooperative group given  $S$ .

(2) The smallest integer  $t$  satisfying (2.2) is called the *threshold of cooperation given group size  $s$* , and is denoted by  $g(s)$ .<sup>4</sup>  $g$  is called the *threshold function of cooperation*.

(3)  $S$  is called an *efficient group* if there exists no  $T \subseteq N$  such that every  $i \in N$  is better off in the action profile  $a^T$  than in the action profile  $a^S$ .

A cooperative group is a group of cooperators in which all members are better off than they would be in the defection equilibrium. A cooperative group  $T$  given  $S$  describes the following situation. A group  $S$  of cooperators prevails while all non-members of  $S$  defect. If the group of cooperators is expanded from  $S$  to  $T$ , then all new members (i.e.,  $T - S$ ) become better off than those free riding on  $S$ . The formation of  $T$  also makes all incumbents (i.e.,  $S$ ) better off since the cooperator's payoff  $u(C, h)$  is increasing in  $h$ . The threshold of cooperation given group size  $s$  is the smallest size of such a cooperative group larger than  $S$ .

From Assumption 2.1, the threshold function  $g$  of cooperation is monotonically increasing in  $s$ , and satisfies  $g(s) \geq s + 2$ . A group  $T$  is a cooperative group given  $S$  if and only if  $t \geq g(s)$ .

**Proposition 2.1.** *Let  $S \subseteq N$ . The following conditions are equivalent.*

(1)  $S$  is an *efficient cooperative group*.

(2)  $S$  is a *maximal cooperative group*.

(3)  $s \geq g(0)$  and  $g(s) > n$ .

*Proof.* (1)  $\Rightarrow$  (2): If  $S$  is not a maximal cooperative group, then there exists some cooperative group  $T$  given  $S$ . By definition, all  $i \in T$  are better off in action profile

---

<sup>4</sup>Without no loss of generality, we assume that  $g(s)$  exists (uniquely) for every integer  $0 \leq s \leq n$ .

$a^T$  than in action profile  $a^S$ . All  $j \notin T$  are also better off, since they free-ride on more cooperators in  $a^T$  than in  $a^S$ .

(2)  $\Rightarrow$  (1): If  $S$  is not an efficient group, then there exists some  $T \subseteq N$  such that every  $i \in N$  is better off in action profile  $a^T$  than in action profile  $a^S$ . If  $T \subset S$ , then every  $i \in T$  is worse off in  $a^T$  than in  $a^S$ , a contradiction. If  $T - S \neq \emptyset$ , then it must be  $u(C, t - 1) > u(D, s)$ . This means that every group of size  $t$  including  $S$  is a cooperative group given  $S$ , contradicting the assumption that  $S$  is a maximal cooperative group.

(2)  $\Leftrightarrow$  (3): This relationship is clear. □

Figure 2.1 illustrates the two payoff functions  $u(C, h)$  and  $u(D, h)$ , and the threshold function  $g$  of cooperation. The horizontal axis represents the number  $h$  of other cooperators. The interval  $XZ$  is the set of cooperative group sizes (except for one member), and the interval  $YZ$  the set of efficient group sizes.

Insert Figure 2.1 here

Let  $G^\infty$  be an infinitely repeated game of the  $n$ -person prisoner's dilemma  $G$ , where each player  $i$  has a common discount factor  $\delta$  ( $0 \leq \delta < 1$ ) for future payoffs and has perfect information on the history of the game. Let  $u_{i,t}$  be player  $i$ 's payoff in every period  $t (= 1, 2, \dots)$ . Player  $i$  maximizes the sum of discounted payoffs,  $\sum_{t=1}^{\infty} \delta^{t-1} u_{i,t}$ . The following well-known result is the starting point of our analysis.

**Proposition 2.2.** *Every cooperative group  $S$  can be sustained as a subgame perfect equilibrium of the repeated game  $G^\infty$  of the  $n$ -person prisoner's dilemma  $G$  if all players' discount factors  $\delta$  satisfy*

$$\delta \geq \frac{u(D, s - 1) - u(C, s - 1)}{u(D, s - 1) - u(D, 0)}. \quad (2.3)$$

This proposition is a special case of the folk theorem in repeated games (for example, see Fudenberg and Maskin 1986). Consider the following trigger strategy for every member in a cooperative group: cooperate first, and keep cooperating as long as all

group members do so, otherwise defect forever. This trigger strategy will be called the *group-trigger strategy* henceforth. Note that punishment may be applied only to group members; non-members are free to defect. If all members of the group are sufficiently patient, then the strategy profile in which they employ the group-trigger strategy and non-members always defect is a subgame perfect equilibrium of  $G^\infty$ . Since the proof of Proposition 2.2 is standard, we omit it. In what follows, we assume that all players are sufficiently patient.

**Assumption 2.2.** Every player  $i$ 's discount factor  $\delta$  satisfies (2.3) for every  $s \geq g(0)$ .

Proposition 2.2 shows a well-known drawback of the folk theorem: a large number of equilibria (cooperative groups). The largest group  $N$  is trivially a cooperative group by Assumption 2.1. Any group is a cooperative group if its size is larger than the threshold of cooperation. In general, an equilibrium of the repeated game  $G^\infty$  involves a conflict between members and non-members since non-members free ride. This fact poses a further question: how such a conflict can be resolved in the repeated game. In other words, which of the cooperative groups will form?

There is another problem with the repeated game  $G^\infty$ : no matter which equilibrium is chosen, it may be vulnerable to the possibility of commitment. For example, consider the largest group  $N$  in equilibrium. Suppose that prior to game play, some player commits himself to not participating in the group  $N$ . Given his decision, all other players may find it beneficial to organize the smaller group  $N \setminus \{i\}$ , which is also sustained in equilibrium as long as it is a cooperative group. The non-participant will be better off as a defector than as a member of the group  $N$ . The same logic may be applied to the smaller group  $N \setminus \{i\}$ , implying a continuing process of members opting out. This argument reveals that a problem with group formation naturally arises in the repeated game  $G^\infty$ . The next section will consider this issue.

### 3 The Model of Group Formation

To consider the problem of group formation in the repeated prisoner's dilemma game  $G^\infty$ , we assume that there exist opportunities for group formation in each period before players take actions. Players attempt to form (and reform) a group of cooperation. The members of a group are bound to implement the group-trigger strategy. The group-trigger strategy is subject to renegotiation. A process of group formation in one period is formulated in two-stages. In the first stage, all non-members of the prevailing group decide independently to participate in the group, or not. In the second stage, both incumbent members and new participants either accept or reject to form a new group, independently. The agreement of the new group is made by unanimity.

Formally, each period  $t (= 1, 2, \dots)$  consists of the following three stages. Let  $S_{t-1} \subset N$  be the group formed in period  $t-1$ , where  $S_0 = \emptyset$ .  $S_{t-1}$  is referred to as the *status-quo group* in period  $t$ .

- (i) Participation stage. All non-members of the status-quo group  $S_{t-1}$  decide independently whether or not to participate in  $S_{t-1}$ . Let  $P_t$  be the set of all new participants.
- (ii) Implementation stage. If the expanded group  $S_{t-1} \cup P_t$  is a cooperative group, then all members of it either accept or reject the new group, independently. The group  $S_t$  is defined by

$$S_t = \begin{cases} S_{t-1} \cup P_t & \text{if all the members accept,} \\ S_{t-1} & \text{otherwise.} \end{cases}$$

In the former case, the new group  $S_t$  forms and its members agree to implement the  $S_t$ -trigger strategy, replacing the (ongoing)  $S_{t-1}$ -trigger strategy. In this case, we say that group  $S_t$  has been implemented. In the latter case,  $S_t$  is rejected and  $S_{t-1}$  remains in existence under the  $S_{t-1}$ -trigger strategy. This rule ( $S_t = S_{t-1}$ ) also applies to the case that the expanded group  $S_{t-1} \cup P_t$  is not a cooperative group.

In this case, the group-trigger strategy for  $S_{t-1} \cup P_t$  can not be a self-binding agreement.

- (iii) Action stage. All players in  $N$  choose their actions. Non-members of  $S_t$  are free to choose their actions, while members of  $S_t$  are bound to cooperate according to the  $S_t$ -trigger strategy.<sup>5</sup>

Let  $\Gamma$  denote the repeated game of the prisoner's dilemma with group formation defined above. As with  $G^\infty$ , every player making a choice in  $\Gamma$  knows the history of all past moves. Every player maximizes the sum of discounted payoffs in all periods. A (pure) strategy  $\sigma_i$  for every player  $i$  in  $\Gamma$  can be defined in the same manner as for  $G^\infty$ . A pure strategy determines player  $i$ 's choice at every possible move in  $\Gamma$ ; in this paper, we do not consider mixed strategies.

A few remarks about the game  $\Gamma$  are in order. First, the group-trigger strategy is employed only if all members of the group agree to it. Each member has veto power. The agreement is self-binding in the sense that it is supported as a subgame perfect equilibrium of the repeated game  $G^\infty$ . When the  $S$ -trigger strategy is agreed upon in period  $t$ , each member receives at least the discounted payoff sum  $\frac{1}{1-\delta}u(C, s-1)$ . If a player defects in the action stage of period  $t$ , he receives at most the discounted payoff sum  $u(D, s-1) + \frac{\delta}{1-\delta}u(D, 0)$ . Thus, if the discount factor satisfies (2.3), then no group member is better off by deviating from  $S$ -trigger strategy.

Second, the members can renegotiate their group-trigger strategy whenever new members wish to participate in the group. If renegotiations fail, then the status-quo group prevails. That is, the threat point of renegotiations is the current agreement of the group-trigger strategy. If all incumbent members and new participants agree, then the group of cooperation is expanded and they will be bound to the new group-trigger strategy.

---

<sup>5</sup>If any member of  $S_t$  defects, then all other members will defect in the next period and forever after according to the  $S_t$ -trigger strategy. The possibility of punishment motivates all members of  $S_t$  to cooperate.

Third, while all non-members are free to choose their actions in every period, this paper will focus on a Markov perfect equilibrium of  $\Gamma$ , in which all non-members of the group always defect (Lemma 4.1). This restriction eliminates the possibility of players cooperating without forming a group. Our treatment may be justified by the view that every cooperative equilibrium of the repeated game is supported by a self-enforcing agreement made in pre-play negotiations. Non-members of the group do not have any means of coordinating to reach such an equilibrium. In real situations, it may be the case that non-members voluntarily cooperate (due to social preferences). However, this paper will show that an efficient level of cooperation *must* be attainable even in the most pessimistic situation where all non-members always defect. We will discuss this issue more in Section 5.

In  $\Gamma$  there exists no initial group, i.e.,  $S_0 = \emptyset$ . The rule of  $\Gamma$  can be easily modified such that an exogenous, non-empty group  $S_0 = S$  is established before the game begins. We denote this game by  $\Gamma(S)$ . A subgame of  $\Gamma$  starting at the beginning of period  $t$  is called a *period  $t$ -subgame*. A period  $t$ -subgame is identical to  $\Gamma(S_{t-1})$ , where  $S_{t-1}$  is the status-quo group in period  $t$ . For a strategy profile  $\sigma = (\sigma_1, \dots, \sigma_n)$  of  $\Gamma$ , the three stages of each period are reduced to strategic-form games (with binary choices) under the assumption that  $\sigma$  will be employed in all future parts of the game. We will call these strategic-form games the *stage games of  $\Gamma$  induced by  $\sigma$* .

For a strategy profile  $\sigma = (\sigma_1, \dots, \sigma_n)$  of  $\Gamma$ , let  $S_t$  be a group formed in period  $t = 1, 2, \dots$  on the play of  $\sigma$ . The sequence  $\{S_t\}_{t=1}^{\infty}$  is called a *group sequence* of  $\sigma$ . By the rule of  $\Gamma$ , a group sequence  $\{S_t\}$  is monotonically increasing, and there exists some integer  $m$  such that  $S_t = S_{t+1}$  for all  $t \geq m$ . Such a group  $S_m$  is called an *absorbing group* of  $\sigma$ . Since  $\{S_t\}$  is monotonically increasing, an absorbing group is unique.

It should be noted that  $\Gamma$  generates all possible plays in the game  $G^\infty$ , since  $\Gamma$  is identical to  $G^\infty$  when no players participate in a group. This implies that the folk theorem also applies to  $\Gamma$ .  $\Gamma$  and  $G^\infty$  have the same sets of subgame perfect equilibrium outcomes. For every cooperative group  $S$ , the following strategy profile is a subgame perfect equilibrium of  $\Gamma$ . Players never participate in any group, and they behave in the

action stage of every period according to  $S$ -trigger strategy. For this reason, we consider the following refinement of a subgame perfect equilibrium of  $\Gamma$ .

**Definition 3.1.** A strategy profile  $\sigma^* = (\sigma_1^*, \dots, \sigma_n^*)$  of  $\Gamma$  is called a *solution* of  $\Gamma$  if it satisfies the following properties.

- (1) (subgame perfection)  $\sigma^*$  is a subgame perfect equilibrium of  $\Gamma$ .
- (2) (Markov property) For every  $i \in N$  and every period  $t$ , the strategy induced by  $\sigma_i^*$  on each period  $t$ -subgame  $\Gamma(S)$  depends only on the status-quo group size  $s$  and on whether or not  $i \in S$ .
- (3) (strictness on play) Let  $\{S_t^*\}_{t=1}^\infty$  be the group sequence of  $\sigma^*$ . Then  $\sigma^*$  prescribes a strict Nash equilibrium on every stage game induced by itself on every period  $t$ -subgame  $\Gamma(S_{t-1}^*)$ .<sup>6</sup>

The definition of the Markov property is standard. It means that every player behaves in the same way in all periods as long as the status-quo group size and his membership status remain unchanged. These factors constitute a payoff-relevant history of the game. A strategy with the Markov property does not depend on the whole history of the game. As an illustration, consider the player set  $N = \{1, 2, 3, 4, 5\}$  and the status-quo groups  $S = \{1, 2\}$  and  $T = \{1, 3\}$ . The Markov property requires only that players 1, 4, and 5 behave in the same way, whether  $S$  or  $T$  is the status-quo group. It is surely true that player 2 and player 3 behave differently under  $S$  and  $T$ , since their memberships in  $S$  and  $T$  differ. However, this similarity does not imply that player 2's strategy under  $T$  is the same as player 3's strategy under  $S$ . That is, the Markov property does not guarantee that every player receives the same discounted payoff sums when the status-quo group sizes are equal. An important implication of the property is that all non-members of a group always defect (Lemma 4.1). It should be remarked that

---

<sup>6</sup>A Nash equilibrium of a strategic-form game is called *strict* if at least one player has a unique best response to it. This definition is weaker than the usual one, in which every player has a unique best response.

our Markov property rules out *inter-period* history dependency of equilibrium strategies, but not *intra-period* history dependency. For example, it does not rule out the possibility that group members' actions during the implementation stage depend upon an outcome of the participation stage. Indeed, this freedom allows them to punish non-participants by not implementing a group (see also the following discussion of property (3)).

If a Nash equilibrium of a stage game in  $\Gamma$  is not strict, then each player has an alternative best response to the equilibrium. The stability of such an equilibrium is weak, since players have no positive incentive to employ their equilibrium strategies. The participation stage game may have a non-strict Nash equilibrium leading to the failure of a new group if the unilateral deviation of each player never affects the outcome. The implementation stage game also has non-strict Nash equilibria due to the unanimity rule. Since group formation requires unanimous acceptance by incumbent and prospective members, all strategy profiles in which at least two members reject a new group are non-strict Nash equilibria. Property (3) requires that a solution should select a strict Nash equilibrium for stage games whenever a status-quo group belongs to the group sequence of the solution.<sup>7</sup> Since every period  $t$ -subgame reached on the solution play must have such a status-quo group, this property means that a solution selects a strict Nash equilibrium for every stage game in all period  $t$ -subgames on equilibrium play. In a period  $t$ -subgame off equilibrium play, the solution may select a non-strict Nash equilibrium for a stage game. The property allows group members to resolve their group as punishments against non-participants. We will discuss this issue more in Section 5.

## 4 Theorem

**Lemma 4.1.** *Let  $\sigma^*$  be a solution of  $\Gamma$ . In  $\sigma^*$ , all non-members of a group defect in every period.*

---

<sup>7</sup>When every member is indifferent to accepting or rejecting the group in the implementation stage, all action profiles are non-strict Nash equilibria. In such a degenerate case, we assume that a group is not implemented.

*Proof.* Since  $\sigma^*$  satisfies the Markov property, the actions of non-members in each period do not affect the play of  $\sigma^*$  in future periods. Given this fact, subgame perfection requires that all non-members should choose defection as their dominant action.  $\square$

**Lemma 4.2.** *Let  $\sigma^*$  be a solution of  $\Gamma$  with a group sequence  $S(\sigma^*) = \{S_t\}_{t=1}^\infty$ . If there exists a cooperative group  $S$  given  $S_{t-1}$ , then  $S_{t-1}$  is expanded to  $S_t$ .<sup>8</sup>*

*Proof.* Let  $F_i(\sigma^*|i \in S_{t-1})$  be player  $i$ 's discounted payoff sum in the period  $t$ -subgame  $\Gamma^t(S_{t-1})$  when  $\sigma^*$  is employed, and let  $i$  be a member of the status-quo group  $S_{t-1}$ . Since  $\sigma^*$  satisfies the Markov property,  $F_i(\sigma^*|i \in S_{t-1})$  depends only on the status-quo group  $S_{t-1}$ . It does not depend on  $t$  or on the entire history of plays before  $\Gamma^t(S_{t-1})$ . Similarly, let  $F_i(\sigma^*|i \notin S_{t-1})$  be player  $i$ 's discounted payoff sum in the period  $t$ -subgame  $\Gamma^t(S_{t-1})$  when  $\sigma^*$  is employed and when  $i$  is a non-member of  $S_{t-1}$ .

By way of contradiction, suppose that  $S_{t-1}$  is not expanded, i.e.,  $S_{t-1} = S_t$ . Since  $\sigma^*$  satisfies the Markov property, it must be true that  $S_{t-1} = S_m$  for all  $m \geq t$ . Then we have

$$\begin{aligned} F_i(\sigma^*|i \in S_{t-1}) &= \frac{1}{1-\delta}u(C, s_{t-1} - 1) \\ F_i(\sigma^*|i \notin S_{t-1}) &= \frac{1}{1-\delta}u(D, s_{t-1}). \end{aligned}$$

The proof is made by three claims. Let  $S_{t-1}$  be the status-quo group in period  $t$ .

Claim 1. In  $\sigma^*$ ,  $S$  is implemented in period  $t$  if it is formed in the participation stage.

Proof of Claim 1. If  $S$  is implemented, then every  $i \in S$  receives the discounted payoff sum

$$u(C, s - 1) + \delta F_i(\sigma^*|i \in S).$$

By the rule of  $\Gamma$ , this value is greater than or equal to  $\frac{1}{1-\delta}u(C, s - 1)$ . On the other hand, if  $S$  is rejected, then every  $i \in S_{t-1}$  receives the discounted payoff sum

$$u(C, s_{t-1} - 1) + \delta F_i(\sigma^*|i \in S_{t-1}) = \frac{1}{1-\delta}u(C, s_{t-1} - 1),$$

---

<sup>8</sup> $S_t$  is not necessarily equal to  $S$ .

and every  $i \notin S_{t-1}$  receives the discounted payoff sum

$$u(D, s_{t-1}) + \delta F_i(\sigma^* | i \notin S_{t-1}) = \frac{1}{1-\delta} u(D, s_{t-1}).$$

Since  $S$  is a cooperative group given  $S_{t-1}$ , it holds that

$$u(C, s-1) > u(D, s_{t-1}) > u(C, s_{t-1}-1). \quad (4.1)$$

The last inequality follows from Assumption 2.1. Relation (4.1) means that the implementation stage in period  $t$  has a strict Nash equilibrium in which  $S$  is agreed. By Definition 3.1, the solution  $\sigma^*$  selects this strict Nash equilibrium. This completes the proof.

Claim 2. There exists some  $T$ ,  $S_{t-1} \subset T \subseteq N$ , such that the following properties hold.

- (i)  $T$  is implemented in period  $t$ .
- (ii) If any  $i \notin T$  participates in  $T$ , then his discounted payoff sum in  $\Gamma^t(S_{t-1})$  weakly decreases.
- (iii) There exists some member  $i \in T - S_{t-1}$  whose discounted payoff sum in  $\Gamma^t(S_{t-1})$  strictly decreases if he does not participate in  $T$ .

Proof of Claim 2. By Claim 1,  $S$  is implemented in period  $t$ . Since  $S$  is a finite set, there exists some  $T_1, S_{t-1} \subset T_1 \subseteq S$  such that  $T_1$  is implemented but that any proper subset of  $T_1$  is not implemented. By construction, if any  $i \in T_1 - S_{t-1}$  does not participate in  $T_1$ , then  $S_{t-1}$  prevails and  $i$  is strictly worse off than he would have been in  $T_1$ . Thus, (iii) holds. If  $T_1$  satisfies (ii), then the claim holds for  $T = T_1$ . If  $T_1$  does not satisfy (ii), then some  $i_1 \notin T_1$  is strictly better off by participating in  $T_1$ . This means that  $T_1 \cup \{i_1\}$  is implemented. We repeat the above arguments by putting  $T_2 = T_1 \cup \{i_1\}$ . Since  $N$  is a finite set, this process reaches some  $T \subseteq N$ . By construction,  $T$  satisfies (i), (ii) and (iii). When  $T = N$ , (ii) is trivially satisfied. This completes the proof.

Claim 3. The group  $T$  in claim 2 is a strict Nash equilibrium in the participation stage of period  $t$ .

Proof of Claim 3. It suffices to show that every  $j \in T - S_{t-1}$  is strictly worse off if he opts out of  $T$  in the participation stage. By (iii) in claim 2, there exists at least one such member  $i \in T - S_{t-1}$ . By the proof of claim 2, two cases are possible: (a) any proper subset of  $T$  is not implemented, and (b)  $T - \{i\}$  is implemented. In case (a), clearly every  $j \in T - S_{t-1}$  is strictly worse off by opting out since  $T - \{j\}$  is not implemented. Consider case (b). In this case  $T - \{i\}$  is implemented, and player  $i$ 's discounted payoff sum in the period  $t$ -subgame  $\Gamma^t(S_{t-1})$  strictly decreases if he opts out of  $T$  during the participation stage. This yields

$$u(C, t - 1) + \delta F_i(\sigma^* | i \in T) > u(D, t - 1) + \delta F_i(\sigma^* | i \notin T - \{i\}). \quad (4.2)$$

Suppose that any player  $j \in T - S_{t-1}$  opts out of  $T$  in the participation stage. If  $T - \{j\}$  is implemented, then  $j$  obtains payoff  $u(D, t - 1)$  in period  $t$ , and thereafter the period  $t + 1$ -subgame  $\Gamma^{t+1}(T - \{j\})$  will be played. By the Markov property of  $\sigma^*$ , all players other than  $i$  and  $j$  behave in the same way in  $\Gamma^{t+1}(T - \{j\})$  as in  $\Gamma^{t+1}(T - \{i\})$ . Also, since  $i$  and  $j$  are group members in  $\Gamma^{t+1}(T - \{j\})$  and  $\Gamma^{t+1}(T - \{i\})$  respectively, their optimal behaviors with respect to other players' strategies with the Markov property are identical in these subgames. Since  $i$  and  $j$  have the same payoff functions, payoff maximization in  $\sigma^*$  implies that  $F_i(\sigma^* | i \notin T - \{i\}) = F_j(\sigma^* | j \notin T - \{j\})$ . Also, the rule of  $\Gamma$  implies that  $F_i(\sigma^* | i \in T) = F_j(\sigma^* | j \in T)$ . By these arguments, it can be seen that (4.2) also holds for  $j$ . Thus,  $j$  is strictly worse off if he opts out of  $T$ . This completes the proof.

By Claim 3, the participation stage of period  $t$  has a strict Nash equilibrium in which all members of  $T - S_{t-1}$  participate in  $S_{t-1}$ . Therefore, the solution  $\sigma^*$  selects a strict Nash equilibrium in the participation stage of period  $t$ , and  $S_{t-1}$  is expanded to  $S_t$ .  $\square$

We are now ready to state the main theorem.

**Theorem 4.1.** *Let  $\Gamma$  be the repeated game of the  $n$ -person prisoner's dilemma with group formation where players are sufficiently patient. There exists a solution  $\sigma^*$  of  $\Gamma$  with an absorbing group  $S^*$  if and only if  $S^*$  is an efficient cooperative group.*

*Proof.* (only-if part) Let  $\sigma^*$  be a solution of  $\Gamma$  with an absorbing group  $S^*$ . It follows from Lemma 4.2 that  $S^*$  must be a maximal cooperative group. Then, by Proposition 2.1,  $S^*$  is an efficient cooperative group.

(if-part) For every efficient cooperative group  $S^*$ , we construct the following strategy profile  $\sigma^*$  of  $\Gamma$ . Let  $S_{t-1}$  be every possible status-quo group in each period  $t$ . Two cases are possible.

(1) When  $S_{t-1} = \emptyset$ ,

- (participation stage) Every  $i \in S^*$  participates in a group, and others do not.
- (implementation stage) If  $S$  has formed in the participation stage, then every  $i \in S$  accepts  $S$  if and only if  $u(C, s - 1) > (1 - \delta)u(D, 0) + \delta u(C, s^* - 1)$ . In particular, when  $S^*$  has formed, every member accepts it.
- (action stage) If any  $S$  is implemented, then all  $i \in S$  cooperate (according to  $S$ -trigger strategy), and all  $j \notin S$  defect.

(2) When  $S_{t-1} \neq \emptyset$ ,

- (participation stage) No  $i \in N - S_{t-1}$  participates in  $S_{t-1}$ .
- (implementation stage) If  $T \supset S_{t-1}$  has formed in the participation stage, then every  $i \in T$  rejects  $T$ .
- (action stage) If any  $S \supset S_{t-1}$  is implemented, then all  $i \in S$  cooperate (according to  $S$ -trigger strategy), and all  $j \notin S$  defect.

When  $\sigma^*$  is employed, the efficient group  $S^*$  is implemented in period 1 and  $S^*$  is not expanded in period 2 and onwards. Thus,  $S^*$  is the absorbing group of  $\sigma^*$ . Clearly,  $\sigma^*$  satisfies the Markov property.

To prove that  $\sigma^*$  is a subgame perfect equilibrium, it suffices to show that every player's choice in  $\sigma^*$  at every decision node is optimal to  $\sigma^*$  itself. First, consider case (1). Clearly, every player's choice in the action stage is optimal. Suppose that a group  $S$  has formed in the participation stage. When  $S$  is implemented in the second stage, every

member  $i \in S$  receives the discounted payoff sum  $\frac{1}{1-\delta}u(C, s-1)$  by the construction of  $\sigma^*$ . When  $S$  is not implemented, he receives  $u(D, 0) + \frac{\delta}{1-\delta}u(C, s^* - 1)$ . Thus,  $\sigma^*$  prescribes every member's optimal choice. Consider now the participation stage. When  $\sigma^*$  is employed, every member  $i \in S^*$  receives the discounted payoff sum  $\frac{1}{1-\delta}u(C, s^* - 1)$ . If  $i$  deviates from  $\sigma^*$ ,  $S^* \setminus \{i\}$  is not implemented since  $u(C, s^* - 2) < (1 - \delta)u(D, 0) + \delta u(C, s^* - 1)$  for sufficiently large  $\delta$ .<sup>9</sup> Thus, every member  $i$  is strictly worse off by deviating from  $\sigma^*$ . If non-member  $j \notin S^*$  participates in  $S^*$ ,  $j$  receives the discounted payoff sum, either  $\frac{1}{1-\delta}u(C, s^*)$  or  $u(D, 0) + \frac{\delta}{1-\delta}u(D, s^*)$ , depending on whether or not  $S^* \cup \{j\}$  is implemented. Either payoff is lower than  $j$ 's discounted payoff sum  $\frac{1}{1-\delta}u(D, s^*)$  for  $\sigma^*$ . Thus, in case (2) it is also clear that every player's choice in  $\sigma^*$  is optimal to  $\sigma^*$ .

Finally, the arguments above show that  $\sigma^*$  induces a strict Nash equilibrium on every period  $t$ -subgame when the status-quo group is empty. Thus,  $\sigma^*$  satisfies the property of strictness on play.  $\square$

The theorem shows that an efficient cooperative group necessarily forms in the repeated prisoner's dilemma when players have the opportunity to form such a group in every period. The possibility of renegotiation allows the group formation device to select an efficient group as the absorbing state of a solution. If the number of reforming status-quo groups is not large enough, however, then an inefficient group may result.

The sufficiency part of the theorem shows only that an efficient cooperative group can be reached as an absorbing group of *some* solution of  $\Gamma$ . The necessity part of the theorem shows that an absorbing group of every solution must be efficient. The next proposition shows the converse, namely that an efficient group is an absorbing group of every solution, provided that it can be reached.

**Proposition 4.1.** *Let  $\sigma^*$  be a solution of  $\Gamma$  with the group sequence  $S(\sigma^*) = \{S_t\}_{t=1}^\infty$ . If  $S_{t-1}$  is an efficient cooperative group, then  $S_{t-1}$  is the absorbing group of  $\sigma^*$ .*

*Proof.* By way of contradiction, suppose that the efficient cooperative group  $S_{t-1}$  is

---

<sup>9</sup>This is the case if  $\delta > \frac{u(C, s^*-2) - u(D, 0)}{u(C, s^*-1) - u(D, 0)}$ .

expanded to  $S_t$  in period  $t$  of  $\sigma^*$ . Then every  $i \in S_t$  receives the discounted payoff sum

$$F_i(\sigma^* | i \in S_{t-1}) = u(C, s_t - 1) + \delta F_i(\sigma^* | i \in S_t), \quad (4.5)$$

in the period  $t$ -subgame  $\Gamma^t(S_{t-1})$ . Since  $S_{t-1}$  is a maximal cooperative group by Proposition 2.1,

$$u(D, s_{t-1}) \geq u(C, n - 1). \quad (4.6)$$

It follows from (4.5) and (4.6) that

$$\frac{1}{1 - \delta} u(D, s_{t-1}) \geq F_i(\sigma^* | i \in S_{t-1}). \quad (4.7)$$

Let  $v$  be the discounted payoff sum that every  $j \in S_t - S_{t-1}$  receives when  $S_t$  is not implemented in period  $t$ . Then it holds that

$$v = u(D, s_{t-1}) + \delta F_i(\sigma^* | i \in S_{t-1}).$$

Note that  $j$  will participate in  $S_t$  in period  $t + 1$  since  $\sigma^*$  has the Markov property. Substituting (4.7) into the equation above yields  $v \geq F_i(\sigma^* | i \in S_{t-1})$ . However, since  $S_t$  is implemented in the play of  $\sigma^*$  by supposition, it must be true that  $F_i(\sigma^* | i \in S_{t-1}) > v$  from (4.5). This is a contradiction.  $\square$

In the proof of Theorem 4.1, we constructed an equilibrium where an efficient group forms immediately. The following proposition shows that there exists a solution of  $\Gamma$  in which an efficient group forms gradually through a sequence of minimum cooperative groups given status-quo groups.

**Proposition 4.2.** *Let  $\Gamma$  be the repeated game of the  $n$ -person prisoner's dilemma with group formation where players are sufficiently patient. Then, there exists a solution  $\sigma^*$  of  $\Gamma$  with a group sequence  $\{S_t\}_{t=1}^{\infty}$  such that  $s_1 = g(0), s_2 = g(s_1), \dots, s_m = g(s_{m-1})$  with  $s_m \leq n < g(s_m)$ , and  $s_t = s_m$  for all  $t \geq m$ .*

*Proof.* Let  $\{S_t\}_{t=1}^{\infty}$  be a group sequence whose sizes satisfy the properties in the proposition. We construct the following strategy profile  $\sigma^*$  of  $\Gamma$  for every period  $t$ . Two cases are possible.

(1) When  $S_{t-1}$  is the status-quo group in period  $t$ ,

- (participation stage) Every  $i \in S_t$  participates in  $S_{t-1}$ , and no others do. In period  $t \geq m + 1$ , no  $i \in N - S_t$  participates in  $S_t$ .
- (implementation stage) If  $S_t$  has formed in the participation stage, then every  $i \in S_t$  accepts  $S_t$ . If  $S \neq S_t$  has formed, then every  $i \in S$  accepts  $S$  if and only if

$$\begin{aligned} & \frac{1}{1-\delta}u(C, s-1) \\ > & u(D, s_{t-1}) + \cdots + \delta^{t'-t+1}u(C, s_{t'}-1) + \cdots + \frac{\delta^{m-t+1}}{1-\delta}u(C, s_m-1) \end{aligned}$$

where  $i \in S_{t'} \setminus S_{t'-1}$  (that is, player  $i$  joins the group  $S_{t'}$  in period  $t' + 1$  for the first time). The RHS represents  $i$ 's discounted payoff sum when  $S_{k-1}$  is implemented in period  $k$  for every  $k \geq t$ .

- (action stage) If any  $S$  is implemented, then all  $i \in S$  cooperate (according to the  $S$ -trigger strategy) and all  $j \notin S$  defect.

(2) When  $S \neq S_{t-1}$  is the status-quo group in period  $t$ ,

- (participation stage) No  $i \in N - S_{t-1}$  participates in  $S$ .
- (implementation stage) If  $T \supset S$  has formed in the participation stage, then every  $i \in T$  rejects  $T$ .
- (action stage) If any  $T \supset S$  is implemented, then all  $i \in T$  cooperate (according to the  $T$ -trigger strategy) and all  $j \notin T$  defect.

When  $\sigma^*$  is employed, each group  $S_t$  ( $1 \leq t \leq m$ ) is implemented in period  $t$ , and  $S_m$  is not expanded in period  $m + 1$  and onwards. Clearly,  $\sigma^*$  satisfies the Markov property.

In analogy with the proof of Theorem 4.1, we will show that every player's choice in  $\sigma^*$  at every decision node is optimal to  $\sigma^*$ . Consider case (1). Clearly, every player's choice in the action stage is optimal. Suppose that  $S_t$  has formed in the participation stage. When  $S_t$  is implemented in the second stage, every member  $i \in S_t$  receives the discounted payoff sum

$$u(C, s_t - 1) + \delta u(C, s_{t+1} - 1) + \cdots + \delta^{m-t}u(C, s_m - 1) + \delta^{m-t+1}u(C, s_m - 1) + \cdots .$$

When  $S_t$  is not implemented, every member  $i \in S_t$  receives at most

$$u(D, s_{t-1}) + \delta u(C, s_t - 1) + \cdots + \delta^{m-t} u(C, s_{m-1} - 1) + \delta^{m-t+1} u(C, s_m - 1) + \cdots .$$

(The first component is replaced with the smaller payoff  $u(C, s_{t-1} - 1)$  if  $i \in S_{t-1}$ ). Since  $u(C, s_t - 1) > u(D, s_{t-1})$ , it is optimal for every  $i \in S_t$  to accept  $S_t$ . Suppose now that a group  $S \neq S_t$  has formed in the participation stage. When  $S$  is implemented in the second stage, every member  $i \in S$  receives the discounted payoff sum  $\frac{1}{1-\delta} u(C, s - 1)$  by the construction of  $\sigma^*$ . When  $S$  is not implemented in the second stage, he receives

$$u(D, s_{t-1}) + \cdots + \delta^{t'-t+1} u(C, s_{t'} - 1) + \cdots + \frac{\delta^{m-t+1}}{1-\delta} u(C, s_m - 1)$$

where  $i \in S_{t'} \setminus S_{t'-1}$ . Thus,  $\sigma^*$  prescribes every member's optimal choice.

Consider now the participation stage. If every member  $i \in S_t$  deviates from  $\sigma^*$ ,  $S_t \setminus \{i\}$  will not be implemented when  $\delta$  is sufficiently large, for the following reason. When  $S_t \setminus \{i\}$  is implemented, every member receives the discounted payoff sum  $\frac{1}{1-\delta} u(C, s_t - 2)$ . If it is not implemented, every member receives the discounted payoff sum

$$u(C, s_{t-1} - 1) + \delta u(C, s_t - 1) + \cdots .$$

When  $\delta$  is sufficiently large, this payoff is larger than  $\frac{1}{1-\delta} u(C, s_t - 2)$ . Therefore, every member  $i \in S_t$  is strictly worse off if they deviate from  $\sigma^*$ . If a non-member  $j \notin S_t$  participates in  $S_t$ ,  $j$  receives the discounted payoff sum, either  $\frac{1}{1-\delta} u(C, s_t)$  or  $u(D, s_{t-1}) + \delta X$ , depending on whether or not  $S_t \cup \{j\}$  is implemented. Here,  $X$  denotes  $j$ 's discounted payoff sum in the period  $t$ -subgame  $\Gamma(S_{t-1})$  when  $\sigma^*$  is employed.  $X$  is given by

$$u(D, s_t) + \cdots + \delta^{t'-t} u(C, s_{t'} - 1) + \cdots + \frac{\delta^{m-t}}{1-\delta} u(C, s_m - 1)$$

when  $j$  joins the group  $S_{t'}$  in period  $t'$  for the first time. Since  $S_t$  is a cooperative group given  $S_{t-1}$ , it can be shown that  $X > u(D, s_{t-1}) + \delta X, \frac{1}{1-\delta} u(C, s_t)$ . Thus, every non-member  $j \notin S_t$  is strictly worse off by participating in  $S_t$ . In case (2), it is clear that every player's choice in  $\sigma^*$  is optimal to  $\sigma^*$ .

Finally, the arguments above show that  $\sigma^*$  induces a strict Nash equilibrium on the period  $t$ -subgame  $\Gamma(S_{t-1})$  for every  $t \leq m$ . Thus,  $\sigma^*$  satisfies the property of strictness on play.  $\square$

## 5 Conclusion

Our theorem shows that when renegotiation is possible, an efficient group of cooperators necessarily forms in the repeated  $n$ -person prisoner's dilemma when individuals are patient. The formation of an efficient group can be either immediate or gradual. To understand the role of renegotiation, it may be helpful to consider an extreme case where there is only one opportunity to negotiate group formation. The first period of this scenario is the same as the first period in  $\Gamma$ , but all future periods have only the action stage. Thus, all group members follow the group-trigger strategy negotiated in the first period (assuming a group formed), and all non-members always defect.<sup>10</sup> If we impose the property of strictness on play, it can be seen that a group is formed in a subgame perfect equilibrium if and only if it is a minimal cooperative group. If we do not, any cooperative group can be formed in a subgame perfect equilibrium. In any case, negotiation may result in an inefficient group. However, if renegotiation is possible, the inefficient group will be expanded to a larger group as Lemma 4.2 shows. In this way, group formation becomes a gradual process that eventually attains efficiency.

The literature contains some other works on gradual cooperation, which show that the level of cooperation (such as tariffs in international trade) increases over time in a two-person repeated prisoner's dilemma. The papers vary in other respects, assuming either adjustment costs (Furusawa and Lai 1999), action irreversibility (Lockwood and Thomas 2002), or uncertainty in the partner's willingness to cooperate (Watson 1999). This work takes a different approach by examining the size of the cooperating group in a multi-player game. However, the more critical difference between our research and related works in the literature is that the latter *assume* the efficiency of an equilibrium. In this paper, efficiency is a result of our approach.

To conclude the paper, we discuss several points in our analysis.

First, we have restricted our attention to the group-trigger strategy as a credible agreement for each group. However, this is just one example of a subgame perfect equilibrium which attains group cooperation in repeated games. Our result holds for any

---

<sup>10</sup>Note that this game is essentially  $\Gamma$  where the players' discount factors are zero.

subgame perfect equilibrium with the same outcome, for example simple strategy profiles (Abreu 1988). In fact, our analysis can be applied to any mechanism which enforces the cooperation of group members. Other examples are the classical Groves-Ledyard mechanism in public goods provision (Groves and Ledyard 1977) and a centralized enforcement institution which penalizes any group member who violates the agreement.

Second, as a refinement of a Markov perfect equilibrium in the repeated game with group formation, we have required that a strict Nash equilibrium be selected for stage games, whenever possible. The selection is applied only for a period-subgame on equilibrium play. In general, the implementation stage game may have two kinds of Nash equilibria: one is a strict equilibrium wherein all members accept a group, and the other is a non-strict equilibrium wherein at least two members reject a group. Our treatment allows group members to punish non-participants by selecting the non-strict equilibrium when the status-quo group is off equilibrium play. It is an empirical issue whether or not participants punish non-participants by resolving their beneficial group. Many experimental studies report that subjects punish opportunistic behavior even if this action means sacrificing their material payoffs. For example, Kosfeld, Okada and Riedl (2009) report experimental observations in a static model of our game  $\Gamma$  in the context of public goods. In four-person games of institution formation where the minimum cooperative group size is two, they find that the implementation rates of two-person and three-person groups are only 37 percent each, while the implementation rate of the maximal group is 90 percent (see Table 2, Kosfeld, Okada and Riedl 2009). Subjects actually punish non-participants in most cases. Kosfeld, Okada and Riedl show that punishing behavior by participants can be explained by the social preference model of Fehr and Schmidt (1999).

Third, while we have considered the group-trigger strategy as a credible equilibrium for repeated games, we have assumed that a cooperating group is renegotiable. In the trigger strategy, if any group member defects, then all other members punish a defector forever. Since such an extreme punishment hurts the punishers, it may be reasonable to allow players to negotiate a new group after the punishment has been

carried out. We have not considered this possibility. In this sense, our approach is unlike that taken in the literature of renegotiation-proof equilibria (see Farrell and Maskin, 1989 and Bernheim and Ray, 1989 among others). In that stream of literature, it is assumed that renegotiation is possible at any time, both ex ante and after each period of play. Renegotiation-proofness has been captured by some cooperative notions such as Pareto domination and internal (and external) consistency. However, it has been shown that a renegotiation-proof equilibrium is not always efficient. In contrast, our approach is purely non-cooperative. We explicitly model renegotiation in group formation and have shown that every absorbing group (which is “renegotiation-proof” by definition) is efficient in a refinement of a subgame perfect equilibrium for a repeated game with group formation. Regarding the possibility of renegotiation, our model is an intermediate case. The standard folk theorem describes one polar case, that only ex ante negotiation is possible. The renegotiation-proof equilibrium describes the other polar case, that negotiation is possible at any time both on and off of the equilibrium path. Our treatment assumes that individuals can renegotiate to form the group of cooperation, but that they can not renegotiate for punishments after defection. A practical message from this paper is that both renegotiation for on-going groups and political commitment not to renegotiate for punishments are effective for attaining efficiency in social dilemma situations.

Finally, our analysis is limited to the repeated prisoner’s dilemma. Extending the proofs to a more general model is left for future work.

## References

- Abreu, D. (1988), “On the theory of infinitely repeated games with discounting,” *Econometrica* 56, 383-396.
- Bernheim, D. B. and D. Ray (1989), “Collective dynamic consistency in repeated games,” *Games and Economic Behavior* 1, 295-326.
- Bloch, F. and A. Gomes (2006), “Contracting with externalities and outside options,”

- Journal of Economic Theory* 127, 172-201.
- Lockwood, B. and J.P. Thomas (2002), "Gradualism and irreversibility," *Review of Economic Studies* 69, 339–356.
- Dixit, A. and M. Olson (2000), "Does voluntary participation undermine the Coase theorem?," *Journal of Public Economics* 76, 309–335.
- Farrell, J. P. and E. Maskin (1989), "Renegotiation in repeated games," *Games and Economic Behavior* 1, 327–360.
- Fehr, E. and K. Schmidt (1999), "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics* 114, 817–868.
- Fudenberg, D. and E. Maskin (1986), "The folk theorem in repeated games with discounting or with incomplete information," *Econometrica* 54, 533–554.
- Furusawa, T. and E.L.-C. Lai (1999), "Adjustment costs and gradual trade liberalization," *Journal of International Economics* 49, 333–361.
- Gomes, A. (2005), "Multilateral contracting with externalities," *Econometrica* 73, 1329-1350.
- Gomes, A. and P. Jehiel (2005), "Dynamic processes of social and economic interaction: on the persistence of inefficiencies," *Journal of Political Economy* 113, 626-667.
- Groves, T, and J. O. Ledyard (1977), "Optimal allocation of public goods: a solution to the 'free rider' problem." *Econometrica* 45, 783-810.
- Hyndman, K. and D. Ray (2007), "Coalition formation with binding agreements," *Review of Economic Studies* 74, 1125-1147.
- Konishi, H. and D. Ray (2003), "Coalition formation as a dynamic process," *Journal of Economic Theory* 110, 1-41.
- Kosfeld, M., A. Okada and A. Riedl (2009), "Institution formation in public goods games," *American Economic Review* 99, 1335-55.
- Okada, A. (2000) The efficiency principle in non-cooperative coalitional bargaining. *Japanese Economic Review* 51, 34-50.
- Schelling, T.C. (1978), *Micro-motives and macro-behavior*, New York: W.W. Norton.

Seidmann, D.J. and E. Winter (1998) A theory of gradual coalition formation. *Review of Economic Studies* 65, 793-815.

Watson, J. (1999), "Starting small and renegotiation", *Journal of Economic Theory* 85, 52-90.

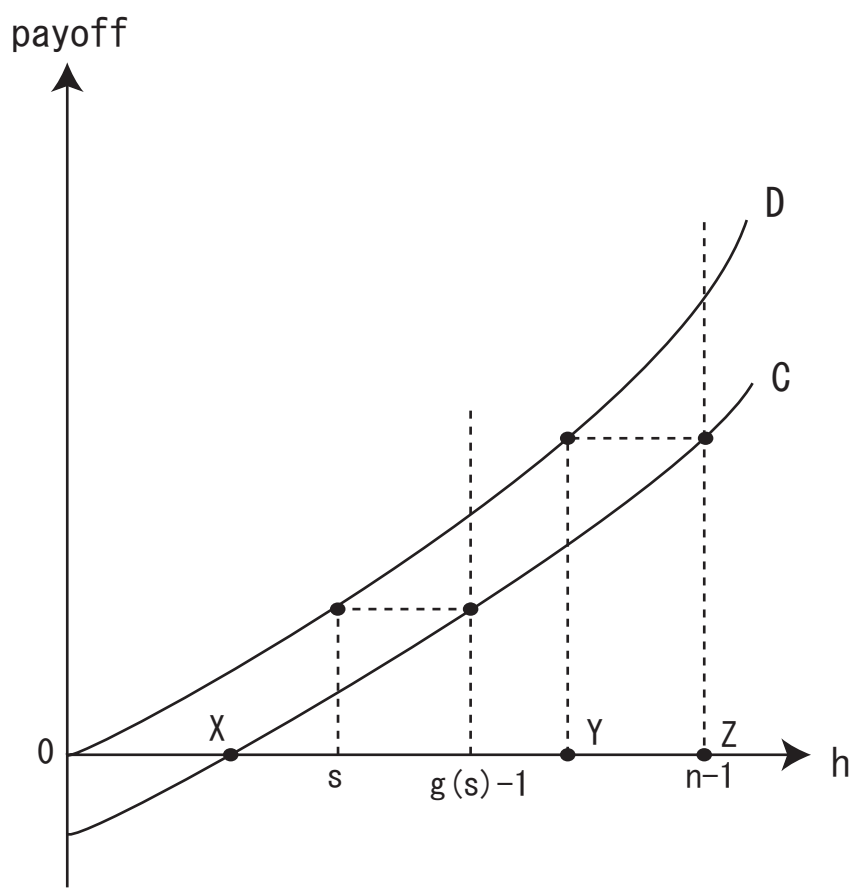


Figure 2.1