Hitotsubashi University
Department of Economics

2-1 Naka, Kunitachi, Tokyo, Japan

Discussion Paper #2014-03

The Formation and Long-run Stability of Cooperative Groups in a Social Dilemma Situation

Toshimasa Maruta and Akira Okada

February 2014
September 2014 (revised)

The Formation and Long-run Stability of Cooperative Groups in a Social Dilemma Situation[1]

Toshimasa Maruta[2] and Akira Okada[3]

September 20, 2014

Abstract: We consider the formation and long-run stability of cooperative groups in a social dilemma situation where the pursuit of individual interests conflicts with the maximization of social welfare. The adaptive play model of Young (1993) is applied to a group formation game where voluntary participants negotiate to create an institution that enforces cooperation. For the class of group formation games with two types, the stochastically stable equilibrium can be characterized in terms of the Nash products of the associated hawk-dove games, which summarize the strategic interaction among the individuals in the game.

Journal of Economic Literature Classification Numbers: C70, C72.

Keywords: Adaptive play, cooperation, evolution, group formation, hawk-dove game, social dilemma, stochastic stability, voluntary participation.

[1]We would like to thank an anonymous referee and Ryoji Sawa for their valuable suggestions and comments.

[2]Advanced Research Institute for the Sciences and Humanities and Population Research Institute, Nihon University, 12-5 Goban-cho, Chiyoda, Tokyo 102-8251, Japan. E-mail:*maruta.toshimasa@nihon-u.ac.jp* Phone: +81 3 5275 9607 Fax: +81 3 5275 9204. This author gratefully acknowledges the financial support from KAKENHI 23530229.

[3]Corresponding author: Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan. E-mail:*aokada@econ.hit-u.ac.jp* Phone: +81 42 580 8599 Fax: +81 42 580 8748

1 Introduction

We consider the formation and long-run stability of cooperative groups in a social dilemma situation where the pursuit of individual interests conflicts with the maximization of social welfare. Public goods provision and common-pool resource management are well-known examples of social dilemmas. The model proposed here captures the fact that individuals in a social dilemma may differ in their willingness to cooperate.

To attain cooperation in a social dilemma, it is critical to implement an appropriate mechanism (i.e., an institution) that prevents selfish individuals from defecting. Such a mechanism is either centralized or decentralized. Centralized institutions observed in reality include police and courts established to attain cooperation or (more generally) social order. Decentralized mechanisms have been well studied in repeated game literature, where the trigger strategy and its variants constitute equilibria in which cooperation is sustained through potential mutual punishments.

An important question in this context is whether and how do individuals voluntarily create such an institution in the first place. Although an enforcement institution is beneficial to all individuals, each individual also has an incentive to free ride on the institution. To consider this problem, two-stage games of group formation are studied (see, e.g., Dixit and Olson 2000 and Kosfeld, Okada, and Riedl 2009). In the first stage of the game, individuals decide independently whether to participate in a group. In the second stage, participants negotiate with each other on establishing an institution. If all participants agree, the institution is established; as a result, all participants cooperate. Meanwhile, non-participants are allowed to free ride on the participants. If a participant does not agree, the institution is not created, and no one cooperates. Kosfeld et al. (2009) show that there exists a subgame-perfect equilibrium in which a number of individuals voluntarily participate in a group, which subsequently establishes an institution. In this manner, a certain level of cooperation can be realized through the two-stage process of group formation. However, there may remain free riders, and hence, the level of cooperation need not to be efficient.

Since the seminal work of Selten (1973) on cartel formation in oligopoly, non-cooperative multi-stage games of group formation with free riders have been applied to various fields such as international environmental agreements (Karp and Simon 2013), R&D spillovers (Katz 1986), and monetary policy (Kohler 2002).

We start with a social dilemma in which individuals differ in their thresholds of cooperation. We then construct a *group formation game*, a strategic game that can be considered a reduced form of the two-stage group formation process previously described. In this game, a participant receives a higher payoff than that in the social dilemma if and only if the number of participants is equal to or larger than the thresholds of any of the participants. This is a necessary (but not sufficient) condition to create an equilibrium cooperative group from the set of participants.

In the first half of this paper, we prove the existence and characterizations of strict Nash equilibria in the group formation game. Because of the asymmetric thresholds, there are generally multiple strict Nash equilibria. There are two distinguishable points on this matter. First, there may be multiple equilibria that differ in the size of the cooperative group. Second, even if a group size is fixed, there may be multiple equilibria that differ in the composition of the members (who have distinct thresholds).

In the second half of the paper, we investigate the equilibrium selection problem that emerges. For this purpose, we employ the adaptive play model of Young (1993), in which the notion of *stochastic stability* is used to identify the equilibrium that is most frequently observed over time in a stochastically perturbed myopic strategy revision process.

In order to focus the selection problem with respect to the size and composition of the equilibrium groups, we simplify by assuming that there is no participation cost. As a result, the strategy profile in which no one participates is a non-strict Nash equilibrium in the group formation game. This feature makes the stochastic stability analysis simpler than usual: An equilibrium cooperative group is stochastically stable if and only if it has the largest resistance to the no-participant equilibrium. To derive explicit selection results, we focus on group formation games in which there are exactly two types of players with respect to their threshold of cooperation. For a class of such games, we show that the stochastically stable equilibrium can be characterized in terms of the *Nash products* of the associated *hawk-dove games*, which represent the best response structure of the game. For another class of games, a selection result follows from the analysis of *unanimity games* developed in Maruta and Okada (2012).

To the best of our knowledge, most existing results on group formation in a social dilemma are static, and few studies have considered its dynamic stability. Myatt and Wallace (2008), however, deserve a special mention. They consider, among others, the dynamic stability of collective action in a threshold public good provision (Palfrey and Rosenthal 1984) under the quantal response strategy revision (McKelvey and Palfrey 1995). In their model, all individuals

have the same threshold of cooperation, which is equal to the minimum number of contributors required to provide a unit of public good. Hence, in any equilibrium with a public good provision, the number of contributors is unique. In other words, there is no multiplicity with respect to the size of the cooperative group. Meanwhile, players in their model differ in values they attach to the public good. In this sense, they address the problem of multiplicity of equilibria with respect to the composition of the members.[1]

The remainder of this paper is organized as follows. Section 2 introduces the group formation game and derives characterizations and the existence of a strict Nash equilibrium. Section 3 reviews adaptive play and the notion of stochastic stability, and then discusses special features in the current setting. Section 4 presents explicit equilibrium selection results for group formation games with two types of players. Appendix formulates a linear programming for mistake counting in the adaptive play, which forms the basis of the formal development of the paper.

2 The Model

2.1 Group formation game

There are n players, with the player set $I = \{1, 2, \cdots, n\}$. Every player $i \in I$ has two actions, $\mathcal{C}$ (cooperation) and $\mathcal{D}$ (defection). Each player i is endowed with a payoff function

$$v^i(a^i, h),\ a^i \in \{\mathcal{C}, \mathcal{D}\},\ h = 0, 1, \cdots, n-1,$$

where a^i is player i's own action and h is the number of others who choose $\mathcal{C}$. We make the following assumptions. For each $i \in I$, the payoff function v^i satisfies:

(**GF1**) $v^i(\mathcal{D}, h) > v^i(\mathcal{C}, h)$ for every $h = 0, 1, \cdots, n-1$

(**GF2**) $v^i(\mathcal{D}, h)$ and $v^i(\mathcal{C}, h)$ are strictly increasing in $h = 0, 1, \cdots, n-1$

(**GF3**) There exists an integer s^i $(2 \leq s^i \leq n)$ such that

$$v^i(\mathcal{C}, s^i - 2) < v^i(\mathcal{D}, 0) < v^i(\mathcal{C}, s^i - 1)$$

[1] From our viewpoint, the model of Myatt and Wallace (2008) can be regarded as a group formation game with a positive participation cost. Hence, the profile in which no one contributes is a strict equilibrium. As a result, they are able to ask (and answer) whether any contribution is more stable than no contribution.

The condition (**GF1**) means that every player has a dominant action $\mathcal{D}$. A player is better off by choosing D than by choosing $\mathcal{C}$ independent of the actions of others. Thus, the game has a unique Nash equilibrium, $(\mathcal{D}, 0)$, in which no players cooperate. It follows from (**GF2**) and (**GF3**) that the Nash equilibrium is Pareto-dominated by the action profile $(\mathcal{C}, n-1)$ in which all players cooperate. The integer s^i in (**GF3**) is the minimum size of a set of cooperators within which player i is better off than in the Nash equilibrium. That is, player i has an incentive to cooperate only if at least $s^i - 1$ others also cooperate. We call s^i the *threshold of cooperation* of player $i \in I$.

We are now ready to define a *group formation game* as follows. The player set is I, each $i \in I$ chooses $\sigma^i \in \{C, D\}$, and the set of action profiles is $\Sigma = \{C, D\}^n$. A subset S of I is called a *successful group* if $|S| \geq s^i$ for every $i \in S$, where $|S|$ is the cardinality of the set S. If all members cooperate in a successful group, they are better off than in the Nash equilibrium. At each $\sigma \in \Sigma$, define $S(\sigma) = \left\{ i \in I \mid \sigma^i = C \right\}$. The payoff $u^i(\sigma)$ for player i at $\sigma \in \Sigma$ is defined as follows, depending on whether $S(\sigma)$ is successful as well as on her own actions.

$$u^i(\sigma) = \begin{cases} v^i(\mathcal{C}, |S(\sigma)| - 1), & \text{if } \sigma^i = C \text{ and } S(\sigma) \text{ is successful,} \\ v^i(\mathcal{D}, |S(\sigma)|), & \text{if } \sigma^i = D \text{ and } S(\sigma) \text{ is successful,} \\ v^i(\mathcal{D}, 0), & \text{if } S(\sigma) \text{ is not successful.} \end{cases} \qquad (\star)$$

The group formation game describes a two-stage process of institution formation discussed in Introduction. In the first stage, every player decides independently whether to participate in an institution. The strategy C means "participation" and D indicates "non-participation." In the second stage, all participants either accept or reject an institution simultaneously. If they all accept it, an institution is created. The institution enforces participants to cooperate. Hence the participants play $\mathcal{C}$. Non-participants are free to choose their actions. Hence the non-participants play $\mathcal{D}$. If an institution is rejected by some participant, the original game is played without an institution. Hence all players choose $\mathcal{D}$. It is easy to show that an institution is created if and only if the set $S(\sigma)$ of participants is successful. The payoff function $(\star)$ is the reduced form of the two-stage process, taking into account the outcome of the second stage. In our setup, an institution can be regarded as a threshold public good in the sense that the number of participants must be larger than their thresholds of cooperation for a successful group. Unlike the standard model (e.g., Palfrey and Rosenthal 1984), the threshold of an institution is not provided exogenously but is determined by participants' incentives to

cooperate.

2.2 Nash equilibrium

We derive characterizations and the existence of strict Nash equilibrium[2] in a group formation game. Let $\sigma = (\sigma^1, \cdots, \sigma^n) \in \Sigma$. As usual, σ^{-i} is the action profile obtained from σ by deleting σ^i. Thus, we can write $\sigma = (\sigma^i, \sigma^{-i})$. For each $\sigma \in \Sigma$, recall that $S(\sigma) = \left\{ i \in I \mid \sigma^i = C \right\}$.

Lemma 1. *Let $\sigma = (\sigma^1, \cdots, \sigma^n) \in \Sigma$.*

(1) *If $\sigma^i = C$ and $S(\sigma)$ is successful but $S(D, \sigma^{-i})$ is not, then $u^i(\sigma) > u^i(D, \sigma^{-i})$.*

(2) *If $\sigma^i = C$ and $S(\sigma)$ is not successful, then $u^i(D, \sigma^{-i}) \geq u^i(\sigma)$.*

(3) *If $\sigma^i = D$ and $S(\sigma)$ is successful, then $u^i(\sigma) > u^i(C, \sigma^{-i})$.*

(4) *$\overline{D} = (D, \ldots, D)$ is a Nash equilibrium in which both C and D are best responses.*

Because the proof of the lemma is straightforward, we omit it. All claims (1)–(4) follow from Assumptions (**GF1**)–(**GF3**), the definition of a successful group, and the equation ($\star$). They reveal aspects of the incentive structure of the group formation game. (1) shows that a player outside the non-successful group has an incentive to join the group if her participation makes the group successful. (2) implies that every member in an unsuccessful group has an incentive to deviate from the group. (3) indicates that players have the incentive to free ride on a successful group whenever possible. The Nash equilibrium in (4) is non-strict. The group formation game may possess many such equilibria. If $S(\sigma)$ is not successful and no unilateral switch leads to a successful group, σ is a non-strict Nash equilibrium.

Let $S \subset I$ be a successful group. A member $i \in S$ is called *critical to* S if $S \setminus \{i\}$ is not successful. S is called *critical* if every $i \in S$ is critical to S.

We can characterize strict Nash equilibria in a group formation game as follows. Let $BR^i(\cdot)$ be the pure best response correspondence of $i \in I$.

Proposition 1. *A strategy profile $\sigma = (\sigma^1, \cdots, \sigma^n)$ in the group formation game is a strict Nash equilibrium if and only if $S(\sigma)$ is successful and critical.*

[2] A strategy profile in a strategic game is a *strict Nash equilibrium* if every strategy is a unique best response to that profile. By "strict equilibrium," we mean a strict Nash equilibrium.

Proof. Let $S(\sigma)$ be successful and critical. Assume that $\sigma^i = D$. Thus, $S(D, \sigma^{-i})$ is successful. Then, by Lemma 1.(3),

$$u^i(D, \sigma^{-i}) > u^i(C, \sigma^{-i}),$$

implying that $BR^i(\sigma) = \{D\}$. Assume next that $\sigma^i = C$. Because $S(\sigma)$ is successful and critical, i is critical to $S(\sigma)$. Hence, $S(\sigma) = S(C, \sigma^{-i})$ is successful, but $S(D, \sigma^{-i})$ is not. By Lemma 1.(1),

$$u^i(C, \sigma^{-i}) > u^i(D, \sigma^{-i}),$$

which means that $BR^i(\sigma) = \{C\}$. Therefore, σ is a strict Nash equilibrium. Conversely, it follows from Lemma 1.(2) and 1.(3) that σ is a strict Nash equilibrium only if $S(\sigma)$ is successful and critical. □

The characterization can be intuitively rephrased in terms of *internal stability* and *external stability*. A group of participants is internally stable if no single member wants to opt out of the group, and it is externally stable if no single outsider wants to join the group. If the group is not successful, it is not internally stable by Lemma 1.(2). When the group is successful, external stability always obtains because anyone outside the group has an incentive to free ride on it. The successful group is internally stable if and only if it is critical. Thus, a strategy profile in the group formation game is a strict Nash equilibrium if and only if the group of participants is both internally and externally stable.

Alternatively, strict Nash equilibria in the group formation game can be characterized in terms of players' thresholds of cooperation. For $S \subset I$ and $k = 2, \cdots, n$, denote by $F_S(k)$ the number of members in S who have threshold of cooperation k. That is, $F_S(k) = \left|\left\{\, i \in S \mid s^i = k \,\right\}\right|$.

Proposition 2. *For a nonempty set S of players, there is a strict Nash equilibrium σ such that $S = S(\sigma)$ if and only if*

$$F_S(2) + \cdots + F_S(|S|) = |S| \quad \textit{and} \quad F_S(|S|) \geq 2.$$

Proof. From the definition of $F_S(\cdot)$, it follows that a group S is successful if and only if $F_S(2) + \cdots + F_S(|S|) = |S|$. Thus, by Proposition 1, it suffices to show that a successful group S is critical if and only if $F_S(|S|) \geq 2$. Suppose that $F_S(|S|) \geq 2$. For every $i \in S$, the group $S \setminus \{i\}$ is not successful because $F_{S\setminus\{i\}}(|S|) \geq 1$. Thus, every member i of S is critical to S. If

$F_S(|S|) = 1$, a unique member i with $s^i = |S|$ is not critical to S because $S \setminus \{i\}$ is a successful group. If $F_S(|S|) = 0$, $s^j \leq |S| - 1$ for every $j \in S$. Therefore, $S \setminus \{j\}$ remains successful. In this case, no member of S is critical. □

Finally, we derive the existence result.

Proposition 3. *Every group formation game possesses a strict Nash equilibrium.*

Proof. Recall that n is the total number of players, and that each threshold s^i is at least two. Thus, there is a number m, $2 \leq m \leq n$, such that $F_I(m) \geq 2$. Let m^* be the largest such number and consider $G = \left\{ i \in I \mid s^i \leq m^* \right\}$. By the choice of m^*, $|I \setminus G| \leq n - m^*$. Thus, $|G| \geq m^*$. Because $F_I(m^*) \geq 2$ and $\left\{ i \in I \mid s^i = m^* \right\} \subset G$, there is a subset $S \subset G$, $|S| = m^*$ that satisfies the condition of Proposition 2. □

3 Adaptive play and Stochastic stability

In this section, we review the stochastic stability approach à la Young (1993) to consider the long-run stability of cooperation in the group formation game. We then present a necessary and sufficient condition for the stochastically stable equilibrium in our setup.

Adaptive play without mistakes is a dynamic adjustment process in discrete time in which a strategic game is played in each period. A state of a period is a sequence of strategy profiles chosen in the last T periods. T is the *memory size* of the process. Each player chooses a best response against her sample, which is a randomly chosen s-length subsequence of the current state where $s \leq T$. s is called the *sample size* of the play. Owing to random sampling, the adaptive play without mistakes is a finite-state Markov chain. A notable property of the chain is that there is a one-to-one correspondence between the set of absorbing states of the chain and the set of strict equilibria in the strategic game. Hence, we may call the absorbing state a strict equilibrium state.

Some noise is then introduced as follows. In each period, a player may fail to choose the best response and end up with a random strategy choice with probability $\epsilon > 0$. If the randomly chosen strategy is not the best response to any sample that might be drawn, the strategy is called a *mistake*. The resulting process is called the *adaptive play with mistakes*, in each period of which a player chooses the best response against a sample with probability at least $1 - \epsilon$ or else makes a mistake. The crucial property of the play with mistakes is that it is irreducible

and aperiodic, and thus, there is a unique stationary distribution μ_ϵ to which the distribution of play converges in the long run. Young (1993) shows that the limit $\mu^* = \lim_{\epsilon \to 0} \mu_\epsilon$ is a stationary distribution of the adaptive play without mistakes. A state is *stochastically stable* if the limiting distribution μ^* puts a positive weight on it. A strict equilibrium in the strategic game is called *stochastically stable* if the corresponding state is stochastically stable.

The notion of *resistance* is the key to identify the stochastically stable state. Because the adaptive play with mistakes is irreducible, there is a positive probability that the play travels from an equilibrium state to another in a finite number of steps. Because any equilibrium state is absorbing in the play with no mistakes, a certain number of mistakes have to occur in an appropriate manner during the transition. The resistance is the minimum number of mistakes that would make that transition possible.

In the current setting, there are two key results. Recall that the sample size and the memory size are s and T, respectively.

Proposition 4. *Consider the adaptive play without mistakes for a group formation game. Assume that $s \leq T/2$. Then, starting from any state, the play reaches a strict equilibrium state in a finite number of steps with positive probability.*

The result ensures that any stochastically stable state corresponds to a strict equilibrium in the group formation game. The proof is provided in Appendix.

For group formation games, the analysis of the adaptive play with mistakes can be drastically simplified. Recall that $\overline{D} = (D, \ldots, D)$, a strategy profile in the group formation game. A state in the adaptive play is called the *$\overline{D}$-state* if its most recent s segment consists entirely of $\overline{D}$. By Lemma 1.(4), any strategy profile can arise as the best response to $\overline{D}$. As a result, if the adaptive play reaches a $\overline{D}$ state, any equilibrium state can follow with no further mistakes. Thus, the following question arises: Starting from an equilibrium state, how and with how many mistakes does the play reach a $\overline{D}$ state? Figure 1 depicts a possible sequence of plays in the adaptive play with mistakes.

Each play in phase 1 is a strict equilibrium, σ, in the group formation game. Phase 1 should be regarded as the most recent s segment of the equilibrium state $E(\sigma)$. In phase 2, every player samples phase 1. While player $m+1$ keep playing the best response, the other players may make mistakes. Specifically, if $X = D$ then it is a mistake for players $1, \ldots, m$. For players $m+2, \ldots, n$, $X = C$ is a mistake. Assume that enough mistakes are made so that

	Phase 1 (s)			Phase 2 (s)			Phase 3 (s)			Phase 4 (s)		
σ^1	C	$\cdots$	C	X	$\cdots$	X	C	$\cdots$	C	D	$\cdots$	D
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
σ^m	C	$\cdots$	C	X	$\cdots$	X	C	$\cdots$	C	D	$\cdots$	D
σ^{m+1}	D	$\cdots$	D	D	$\cdots$	D	C^{br}	$\cdots$	C^{br}	D	$\cdots$	D
σ^{m+2}	D	$\cdots$	D	X	$\cdots$	X	D	$\cdots$	D	D	$\cdots$	D
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
σ^n	D	$\cdots$	D	X	$\cdots$	X	D	$\cdots$	D	D	$\cdots$	D

Figure 1: A state transition in the adaptive play with mistakes.

player $m+1$ can choose C as a best response to phase 2. Consider the sample assignment in which player $m+1$ samples phase 2 and all others sample phase 1. Then, phase 3 arises exactly as depicted, and such sample assignment is possible if $s \leq T/3$. Denote the profile in phase 3 by $\tilde{\sigma} = (\tilde{\sigma}^{m+1}, \sigma^{-(m+1)})$, where $\tilde{\sigma}^{m+1} = C$. $\tilde{\sigma}$ cannot be a strict equilibrium, but $S(\tilde{\sigma})$ might be successful. Even if it is successful, no player in $\{1, \ldots, m\} = S(\sigma)$ is critical to it unless $s^{m+1} = m+1$.

In phase 4, let players $1, \ldots, m$ best respond to phase 3. Assuming that $s^{m+1} \neq m+1$, they choose D. Meanwhile, let players $m+1, \ldots, n$ best respond to the last available segment of phase 1 and the most recent realizations of phase 4. Then their choice is also D. Hence, a $\overline{D}$ state is reached, from which any state can follow.

To summarize and generalize the observation, we name the relevant objects. In any state transition from $E(\sigma)$ to another equilibrium state, there is the *earliest* period in which at least one player chooses a strategy $\tilde{\sigma}^i \neq \sigma^i$ as a *best response* for the *first time*. We call such a player the *first exitor* and the resulting profile, $\tilde{\sigma} = (\tilde{\sigma}^i, \sigma^{-i})$, the *first exit*.[3] In Figure 1, the profile in phase 3 is the first exit, $\tilde{\sigma}^i = C$, and the first exitor is $i = m+1$. Another type of first exit exists where the first exitor is a cooperator[4] in the original equilibrium, and thus, $\tilde{\sigma}^i = D$. In this case, the set $S(\tilde{\sigma})$ of cooperators in the first exit is not successful. One can verify that with appropriate sample assignments a $\overline{D}$ state can arise with no extra assumption. In general,

[3] In general, there may be several first exitors, but we can focus on the single first exitor case because we are searching for the minimum number of mistakes that induces a first exit.

[4] Given a strict equilibrium σ, a *cooperator* is a player i such that $\sigma^i = C$. A *free rider* is a player j such that $\sigma^j = D$.

the number of mistakes required for a first exit depends on the type of the first exitor and her strategy in the original equilibrium. We have observed that with enough mistakes that induce a first exit, any equilibrium state can be reached, via $\overline{D}$ state, with no extra mistakes under the assumptions that no player in $S(\sigma)$ is critical to $S(\tilde{\sigma})$ and $s \leq T/3$. We state this observation as a result. For each type of first exit from a given original equilibrium state $E(\sigma)$, evaluate the minimally possible number of mistakes required to realize it. Then, minimize this number with respect to the type of first exit. The resulting number is called the *exit resistance* of σ.[5] This is the minimally possible number of mistakes to leave σ. Given a group formation game G, a positive integer m is an *equilibrium size* if G has a strict equilibrium σ such that $|S(\sigma)| = m$. *Thresholds are disconnected* in G if for every equilibrium size m, there is no $i \in I$ such that $s^i = m + 1$.

Proposition 5. *Consider the adaptive play with mistakes for a group formation game. Assume that $s \leq T/3$ and that thresholds are disconnected in G. Then, the resistance from an equilibrium state E to another equilibrium state E' is equal to the exit resistance of E. Furthermore, an equilibrium is stochastically stable if and only if the corresponding state has the maximum exit resistance.*

The last part of the proposition can be proven using the general minimum tree argument of Young (1993, Theorem 4) together with the facts that a $\overline{D}$ state can be reached with the minimally possible number of mistakes to leave the original equilibrium state, and that any equilibrium state can be reached from a $\overline{D}$ state with no extra mistake. In Appendix, we formulate a linear program especially designed to pin down the exit resistance.

4 Stochastic stability in games with two types

To derive some explicit equilibrium selection results, we restrict the analysis to a special case where there are only two types of players with respect to the threshold of cooperation.

4.1 Group formation game with two types

Let I be the player set of a group formation game where $|I| = n$. The game is called a *group formation game with two types* if there are natural numbers $2 \leq m < M \leq n$ such

[5] A precise definition is provided in Appendix.

that $s^i \in \{m, M\}$ for every $i \in I$, $|I_m| \geq m$, $|I_M| \geq 2$, where $I_m = \left\{ i \in I \mid s^i = m \right\}$ and $I_M = \left\{ i \in I \mid s^i = M \right\}$. A player is *type* m if $s^i = m$ and *type* M if $s^i = M$. We assume that the same type of players have identical payoff function.[6]

By Proposition 2, a group formation game with two types has exactly two equilibrium group sizes, m and M, and each M-sized equilibrium contains at least two type-M cooperators. Furthermore, equilibria in this class of games exhibit great variety depending on m, M, and the number of players in each type. For example, consider an M-sized equilibrium. If $M < n$, there may or may not be a type-M free rider, a type-m free rider, or a type-m cooperator. In addition, there are multiple M-sized equilibria that differ in their composition. In contrast, there is a unique M-sized equilibrium if $M = n$. Similarly, in an m-sized equilibrium type-m cooperators and type-m free riders coexist if and only if $m < |I_m|$. In this section, we consider the case that $m < |I_m|$ and $M < n$ and the case that $m = |I_m|$ and $M = n$. In the first case, we associate a pair of *hawk-dove games* with the group formation game, and their Nash products are shown to characterize the stochastically stable equilibrium. In the latter case, a selection result follows from the analysis of *unanimity games* developed in Maruta and Okada (2012).

For the remainder of the analysis, let G be a group formation game with two types. We employ the following notations. We designate a generic type by $\tau \in \{m, M\}$. Denote by C_k^τ the group formation game payoff of a type-τ player when she plays C as a member of a successful group of size $k+1$. Denote by D_k^τ the group formation game payoff of a type-τ player when she plays D as a free rider on a successful group of size k. Denote the payoff at $\overline{D}$ by D_0^τ. Note that D_k^τ or C_k^τ may be ill-defined for some k. For example, D_{m-1}^m is always ill-defined. For some values, this definition depends on the threshold distribution in G. For example, C_m^m is well-defined if and only if $m < |I_m|$. In words, D_k^τ and C_k^τ denote the payoff values in the underlying social dilemma (see (**GF1**)–(**GF3**)) that may result as the payoff in G.

4.2 Hawk-dove games

Hawk-dove games and their Nash products (Harsanyi and Selten 1988) play central roles in this subsection. See Figure 2. For $\tau \in \{r, c\}$, let $d^\tau > h^\tau$. A 2×2 game in the figure is a *hawk-dove* (or *chicken*) *game* if $\alpha^\tau > 0$ and $\beta^\tau > 0$. It has two strict equilibria, $(Hawk, Dove)$ and $(Dove, Hawk)$. The number $\left(\frac{\alpha^r}{\alpha^r+\beta^r}\right)\left(\frac{\beta^c}{\alpha^c+\beta^c}\right)$ is called the (normalized) *Nash product* of

[6] In general, players with an identical threshold may attach different payoff values to successful cooperation in a strict equilibrium.

	Dove	*Hawk*
Dove	d^r, d^c	$\beta^r + h^r, \alpha^c + d^c$
Hawk	$\alpha^r + d^r, \beta^c + h^c$	h^r, h^c

Figure 2: A hawk-dove game.

$(Hawk, Dove)$. The Nash product of $(Dove, Hawk)$ is $\left(\frac{\beta^r}{\alpha^r+\beta^r}\right)\left(\frac{\alpha^c}{\alpha^c+\beta^c}\right)$. In this subsection, we assume that G satisfies the following condition.

(**HD**) $m < |I_m| < |I_m| + 2 < M < n \quad \text{and} \quad |I_M| < M.$

This is the condition that determines the class of games in which stochastically stable equilibrium is characterized by Nash products of the associated hawk-dove games. Under (**HD**), thresholds are disconnected. Hence Proposition 5 is applicable to G.

Because $m < |I_m|$, every size m equilibrium in G contains a type-m free rider. Because $M < n$, every M-sized equilibrium contains either a type-m or type-M free rider, or both. Because $|I_M| < M$, every M-sized equilibrium contains both type-m and type-M cooperators, and there is an (Mm, m) *equilibrium*, an M-sized equilibrium in which all free riders are of type m. There is an (Mm, M) *equilibrium*, an M-sized equilibrium in which all free riders are of type M, if and only if $|I_m| + 2 \leq M$.[7] We assume that $|I_m| + 2 < M$, which implies that there are at least *three* type-M cooperators in every M-sized equilibrium.[8]

An m-sized equilibrium is unique up to a permutation of the players that preserves their types. Moreover, the same is true for the (Mm, M) and (Mm, m) equilibria. Hence, up to a type-preserving player permutation, our problem has been reduced to an equilibrium selection among *three* strict equilibria. In order to determine which of the three is stochastically stable, we introduce two auxiliary 2×2 games. Consider the games in Figure 3. By the definition of

[7]There are (Mm, Mm) equilibria, but their resistance is weakly bounded from above by that of the (Mm, M) equilibrium. Thus, the (Mm, Mm) equilibrium can be stochastically stable only if the (Mm, M) equilibrium is stochastically stable. Both are M-sized equilibria, in which there are type-M free riders. We exclude the (Mm, Mm) equilibrium from the analysis because it is similar enough to the (Mm, M) equilibrium.

[8]We assume $|I_m| + 2 < M$ to avoid tedious case distinctions. See the penultimate paragraph in the proof of Lemma A1 in Appendix. Other implications of (**HD**) are $|I_m| \geq 3$, $|I_M| \geq 4$, $M \geq 6$, $n \geq 7$, and $M \geq m + 4$. We can "construct" a threshold distribution that satisfies the condition as follows. Choose $n \geq 7$. Then, choose $m, M \in \{2, \ldots, n-1\}$ such that $M \geq m + 4$. Finally, choose $|I_m| \in \{m+1, \ldots, M-3\}$. Note that $\{m+1, \ldots, M-3\} \neq \varnothing$ since $M \geq m + 4$.

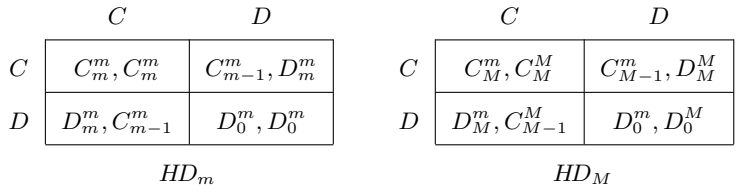

Figure 3: 2×2 hawk-dove games.

the group formation game, both are hawk-dove games. Consider an m-sized equilibrium in G. Pick a type-m cooperator and a type-m free rider. Fixing the behaviors of all others, the game HD_m describes the strategic interaction of these two players. Similarly, consider an M-sized equilibrium and choose a cooperator and a free rider whose types are different,[9] and fix the behaviors of all others. Setting the type-m player as the row player and the type-M player as the column player, the game HD_M describes the strategic interaction of the two players.

There are three Nash products to consider. The Nash product Π_m of HD_m,[10] $\Pi_{(C,D)}$ of (C, D) in HD_M, and $\Pi_{(D,C)}$ of (D, C) in HD_M. Setting

$$\mu_m^m = \frac{C_{m-1}^m - D_0^m}{C_{m-1}^m - D_0^m + D_m^m - C_m^m}, \quad \mu_M^m = \frac{C_{M-1}^M - D_0^M}{C_{M-1}^M - D_0^M + D_M^M - C_M^M}, \quad \mu_M^M = \frac{C_{M-1}^m - D_0^m}{C_{M-1}^m - D_0^m + D_M^m - C_M^m},$$

these Nash products are given as follows:

$$\Pi_m = \mu_m^m(1 - \mu_m^m), \quad \Pi_{(C,D)} = \mu_M^M(1 - \mu_M^m), \quad \Pi_{(D,C)} = \mu_M^m(1 - \mu_M^M).$$

In addition, we introduce the *Nash product of* HD_M, denoted by Π_M, as the number

$$\Pi_M = \mu_M^{\tau^*}\left(1 - \mu_M^{\tau^*}\right),$$

where $\mu_M^{\tau^*} = \min\{\mu_M^m, \mu_M^M\}$. This is the Nash product of the symmetric hawk-dove game, which is analogous to HD_m but played by two players of the same type in an M-sized equilibrium.[11]

[9] If the equilibrium in question contains exactly three cooperators, then it is an (Mm, M) equilibrium; thus, a type-m cooperator and a type-M free rider are chosen to play HD_M. If it were the case that $|I_m| + 1 \geq M$, then there would be an M-sized equilibrium that contained exactly two type-M cooperators and a type-m free rider. If we focused on a type M cooperator and a type m free rider, then the resulting 2×2 game would not be a hawk-dove game.

[10] The two strict equilibria in a symmetric hawk-dove game are indistinguishable in terms of the Nash product. Identifying the two products, we say *the* Nash product of the symmetric hawk-dove game.

[11] If $\mu_M^m \leq \mu_M^M$, the relevant symmetric game should be the 2×2 hawk-dove game played by two type-M players.

The main result is proven under another set of assumptions. In evaluating exit resistances, they ensures that we only need to look at the sequences of plays during which two or more mistakes never occur simultaneously. See Appendix for details. For each $\tau \in \{m, M\}$ and $m \leq l \leq n-1$ where both D_l^τ and C_l^τ are well-defined, set $\Delta^\tau(l) = D_l^\tau - C_l^\tau$, which we call the *type-τ incentive to free ride on l* cooperators. This represents the payoff increase for a type-τ player when she switches from C to D when the remaining l cooperators create a successful group by themselves.

(**P1**) For every $k \geq 1$ such that $m+1 \leq m+k \leq n-1$, $\Delta^m(m+k)/\Delta^m(m) \leq (k+1)$. For every k (possibly negative) such that $m \leq M+k \leq M-4$ or $M+1 \leq M+k \leq n-1$, $\Delta^m(M+k)/\Delta^m(M) \leq |k+1|$.

(**P2**) For every $k \geq 1$ such that $M+1 \leq M+k \leq n-1$, $\Delta^M(M+k)/\Delta^M(M) \leq (k+1)$.

(**P3**) $\frac{C^m(M-1)-D^m(0)+\Delta^m(m)}{C^m(m-1)-D^m(0)+\Delta^m(m)} \leq M-m-1$.

(**P4**) $C^M(M-1) - D^M(0) \leq (M-m-2)(D^M(m) - D^M(0))$.

The first condition in (**P1**) means that the ratio of a type-m player's incentive to free ride on $m+k$ to that on m is bounded above by $k+1$. Other conditions in (**P1**) and (**P2**) are interpreted similarly. Note that (**P1**) and (**P2**) hold if $\Delta^\tau(l)$ is constant in l. (**P3**) requires that the cooperative payoff $C^m(M-1)$ of type m in an M-sized equilibrium is not extremely large. (**P4**) requires that the cooperative payoff $C^M(M-1)$ of type M in an M-sized equilibrium is not extremely large relative to the free-riding payoff $D^M(m)$ of that type in an m-sized equilibrium.

Proposition 6. *Under assumptions* (**HD**) *and* (**P1**)–(**P4**)*, m-sized equilibria are stochastically stable if and only if $\Pi_m \geq \Pi_M$. Furthermore, (Mm, M) equilibria are stochastically stable if and only if $\Pi_M \geq \Pi_m$ and $\mu_M^M \geq \mu_M^m$, and (Mm, m) equilibria are stochastically stable if and only if $\Pi_M \geq \Pi_m$ and $\mu_M^m \geq \mu_M^M$.*

The proof is provided in Appendix. The following discussion reveals the intuition of the result. In Figure 4,[12] the Nash products in HD_M are depicted. The Nash product $\Pi_{(C,D)}$ is the area of the southeast rectangle, and it is a proxy of the resistance of (Mm, M) equilibrium.

[12]The horizontal axis measures the probability that type m chooses C. The vertical axis measures the probability that type M chooses C. The kinked bold lines denote the best responses of each player.

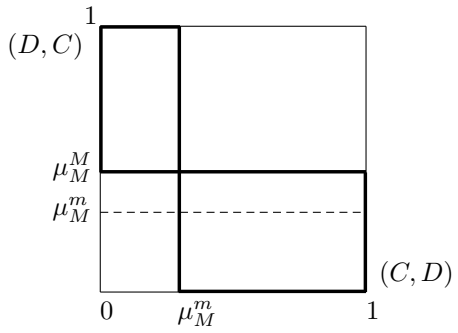

Figure 4: Best response correspondences in HD_M.

The width of the rectangle, $1 - \mu_M^m$, measures the difficulty for type-M players to switch from D to C. Similarly, the height of the rectangle, μ_M^M, measures the difficulty for type-m players to switch from C to D. It is clear that $\Pi_{(C,D)} \geq \Pi_{(D,C)}$ if and only if $\mu_M^M \geq \mu_M^m$, in which case the resistance of the (Mm, M) equilibria can be shown to be equal to or larger than that of the (Mm, m) equilibria, and Figure 4 shows such a case. The Nash product $\Pi_{(C,D)}$, however, should not be compared to that of HD_m directly, because it overestimates the resistance of the (Mm, M) equilibria. Recall that according to (**HD**), every M-sized equilibrium contains both type-m and type-M cooperators. Hence, the difficulty for a type-M player to switch from C to D, which is measured by μ_M^m, should also be considered. Therefore, it is the product $\mu_M^m(1 - \mu_M^m)$, as opposed to $\Pi_{(C,D)} = \mu_M^M(1 - \mu_M^m)$, that should be compared to the Nash product of HD_m. Recalling its definition, the product is nothing but HD_M, which is the area of the southeast rectangular trimmed by the dashed line.

In the following example, m, M, $|I_m|$, and $|I_M|$ are assumed to satisfy (**HD**).

Example 1. Consider a voluntary contribution game of a linear public good. There are n players, each of whom initially owns one unit of a private good. Each player i decides whether to contribute one unit of the private good ($\sigma^i = 1$) or not ($\sigma^i = 0$). For a strategy profile $\sigma \in \{0, 1\}^n$, the payoff $v^i(\sigma)$ of player i is given by $v^i(\sigma) = a^i \sum_{j \in I} \sigma^j - \sigma^i$, where $0 < a^i < 1$ is the marginal benefit of the contribution of player i. Assume that $a^i \in \{a_M, a_m\}$ and that

$$m - 1 \leq \frac{1}{a_m} < m < M - 1 \leq \frac{1}{a_M} < M,$$

so that the threshold of a type-$\tau \in \{m, M\}$ player is τ. If $a_m \leq 1/2$, then $m \geq 3$, and the assumptions (**P1**)–(**P4**) are satisfied. We can then verify that $\mu_m^m = (a_m m - 1)/a_m(m - 1)$, $\mu_M^m = (a_M M - 1)/a_M(M - 1)$, and $\mu_M^M = (a_m M - 1)/a_m(M - 1)$. Because $a_M < a_m$, $\mu_M^m < \mu_M^M$.

Thus, (Mm, m) equilibria are *not* stochastically stable. We can also verify that $\mu_m^m \leq 1/2$ and $\mu_M^m \leq 1/2$. Thus,[13] the (Mm, M) equilibria are stochastically stable if and only if $\mu_M^m \geq \mu_m^m$ or

$$\frac{a_M M - 1}{a_M(M-1)} \geq \frac{a_m m - 1}{a_m(m-1)}.$$

The left-hand side indicates how much payoff a type-M cooperator receives when the number of cooperators is M relative to the free-riding payoff when the number of cooperators is $M - 1$.[14] The right-hand side indicates a similar incentive of a type-m cooperator in an m-sized equilibrium. The result shows that an (Mm, M) equilibrium is stochastically stable if the type-M player's incentive to cooperate (in the sense indicated above) is higher than that of the type-m player.

4.3 Unanimity games

In this subsection, we assume that G satisfies the following condition.

(**U**) $m = |I_m| < M = n$.

There is a group of players who are the most reluctant to cooperate, and they are motivated to cooperate only if all others do. Thus, their threshold is n. Meanwhile, some players are less reluctant to cooperate, and their threshold is $m < n$. It is assumed that the number of such players is exactly m. Under (**U**), G has exactly two equilibria. In the m-sized equilibrium σ_m, $S(\sigma_m) = I_m$. In the n-sized equilibrium σ_n, $S(\sigma_n) = I_m \cup I_n = I$, that is, full cooperation. No other strategy profile contains a successful group.[15] It follows that for every type-m player, there is no strategy profile at which D pays off higher than C does. Therefore, C (weakly) dominates D for type-m players, and we can safely fix their strategies as C.[16] The resulting game is an $(n - m)$-person *unanimity game*, played only be type-n players. In this game,

[13] For any real numbers $p, q \in (0, 1)$, $p(1-p) \geq q(1-q)$ if and only if $\min\{p, 1-p\} \geq \min\{q, 1-q\}$.

[14] The numerator is equal to the payoff in the (Mm, M) equilibrium, but the denominator never results as the payoff in the group formation game because $M - 1$ players never achieve a successful group. However, in this particular example, $C_{M-1}^M + D_0^M + D_M^M - C_M^M = v^M(\mathcal{D}, M-1)$.

[15] *Proof.* Pick $\sigma \neq \sigma_n$ and assume that $S(\sigma)$ is successful. Because $|S(\sigma)| < n$, no player in $S(\sigma)$ has threshold n. Thus, $S(\sigma) \subset I_m$, and therefore, $|S(\sigma)| \leq m$. Because $S(\sigma)$ is successful, $|S(\sigma)| = m$. Therefore, $\sigma = \sigma_m$.

[16] It can be shown that for each sample that induces a first exit by a type-m player, there is a sample that induces a first exit by a type-n player such that the number of mistakes in the latter is no larger than that in the former.

if all players choose D, the m-sized equilibrium arises. If all players choose C, the n-sized equilibrium arises. Any other strategy profile yields an unsuccessful payoff.

Denote

$$r(m) = \frac{D_m^n - D_0^n}{C_{n-1}^n - D_0^n + D_m^n - D_0^n}, \quad r(n) = \frac{C_{n-1}^n - D_0^n}{C_{n-1}^n - D_0^n + D_m^n - D_0^n}.$$

The next result follows from Maruta and Okada (2012, Proposition 4).

Proposition 7. *Assume* (**U**). *If* $\min\{r(m), r(n)\} < \frac{1}{n-m-1}$, *then the* k^**-sized equilibrium in* G *is stochastically stable, where* $r(k^*) = \max\{r(m), r(n)\}$. *Otherwise, both equilibria are stochastically stable.*

Considering only type-n players, either equilibrium Pareto-dominates the other, depending on $C_{n-1}^n > D_m^n$ or vice versa. The result states that only the Pareto-dominant equilibrium can be uniquely stochastically stable, and it is stochastically stable only if it Pareto-dominates the other in a sufficiently wide margin. How wide should be the margin? If $C_{n-1}^n > D_m^n$, a full-cooperation equilibrium is uniquely stochastically stable only if $C_{n-1}^n - D_0^n > (n - m - 2)(D_m^n - D_0^n)$. On the one hand, the unique selection always occurs if $n - m$ equals two or three. On the other hand, the unique selection becomes harder to obtain as $n - m$ increases beyond three.

5 Conclusion

We considered characterizations, existence, and the stochastic stability of equilibrium cooperative groups that may arise from the group formation game, which modeled a process of institution formation in a social dilemma. In the group formation game, the enforcing institution would be formed if an agreement among the participants was reached. In this study, it was assumed that individuals could negotiate with each other to reach such an agreement with no associated cost. Although this is an idealized situation, the problem of institution formation is far from trivial even in this setting, because a serious equilibrium selection problem remains. When individuals differ in their willingness to cooperate, multiple equilibrium cooperative groups might exist that differ both in size and composition of their members. To resolve this problem, we employed the notion of stochastic stability. For classes of group formation games with two types, we were able to capture the essence of the strategic interaction among the individuals through the associated hawk-dove games or the associated unanimity games, in

terms of which a stochastically stable equilibrium can be characterized. These results should help us compare relative stability among multiple equilibrium cooperative groups.

References

Dixit, A. and M. Olson (2000), "Does Voluntary Participation Undermine the Coase Theorem?" *Journal of Public Economics* **76**, 309–335.

Harsanyi, J.C. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*, Cambridge: MIT Press.

Karp, L. and L. Simon (2013), "Participation Games and International Environmental Agreements: A non-parametric model," *Journal of Environmental Economics and Management* **65**, 326–344.

Katz, M.L. (1986), "An Analysis of Cooperative Research and Development," *The RAND Journal of Economics* **17**, 527–543.

Kohler, M. (2002), "Coalition Formation in International Monetary Policy Games," *Journal of International Economics* **56**, 371–385.

Kosfeld, M., A. Okada, and A. Riedl (2009), "Institution Formation in Public Goods Games," *American Economic Review* **99**, 1335–1355.

Maruta, T. and A. Okada (2012), "Stochastically Stable Equilibria in n-person Binary Coordination Games," *Mathematical Social Sciences* **63**, 31–42.

McKelvey, R.D. and T.R. Palfrey (1995), "Quantal Response Equilibria for Normal Form Games,E' *Games and Economic Behavior* **10**, 6–38.

Myatt, D. and C. Wallace (2008), "Does One Bad Apple Spoil the Barrel? An Evolutionary Analysis of Collective Action," *Review of Economic Studies* **75**, 499–527.

Palfrey, T. and H. Rosenthal (1984), "Participation and the Provision of Discrete Public Goods: A Strategic Analysis," *Journal of Public Economics* **24**, 171–193.

Selten, R. (1973), "A Simple Model of Imperfect Competition, where 4 Are Few and 6 Are Many," *International Journal of Game Theory* **2**, 141-201.

Young, P.H. (1993), "The Evolution of Conventions," *Econometrica* **61**, 57–84.

Appendix

Proof of Proposition 4

A strategy profile $\sigma \in \Sigma$ in the group formation game is called *successful* if $S(\sigma)$ is successful.

Proof of Proposition 4. First, we show the result under the assumption that $(s,T) = (1,2)$. Recall that $BR^i(\cdot)$ is the pure best response correspondence. Set $BR = \Pi_i BR^i$. Pick $\sigma_0 \in \Sigma$ and consider the sequence $\langle \sigma_0, \sigma_1, \ldots, \sigma_m \rangle$ such that $\sigma_{k+1} \in BR(\sigma_k)$ and $m = |S(\sigma_1)|$, with the convention that if $D \in BR^i(\sigma)$, D is taken. We construct the desired sequence by modifying it if necessary. If there is $k \geq 1$ such that σ_k is not successful, let players in $S(\sigma_k)$ best respond to σ_k and let the others best respond to σ_{k-1}. Then, $\overline{D}$ arises by Lemma 1.(2). By Lemma 1.(4), any strict equilibrium can arise as a best response to $\overline{D}$. Next, assume that every σ_k, $k \geq 1$, is successful. If every σ_k is successful but not critical, it follows from Lemma 1.(3) that $|S(\sigma_{k+1})| < |S(\sigma_k)|$, which in turn implies that $|S(\sigma_m)| \leq 1$ because $m = |S(\sigma_1)|$. However, σ_m cannot be successful if $|S(\sigma_m)| \leq 1$. Therefore, we must conclude that there is a σ_k that is successful and critical. By Proposition 1, this σ_k is a strict equilibrium. We have shown the desired result for $(s,T) = (1,2)$. It remains to translate the sequence into state transitions in the adaptive play with any (s,T) such that $s \leq T/2$. This is straightforward and we omit the details. □

Computing the exit resistance

Proposition 5 states that a stochastically stable equilibrium in a group formation game can be found by comparing exit resistances. In order to find it explicitly, we need to compute the values of exit resistances. Let us introduce the linear program that works for this purpose.

Consider a group formation game. The set of strategy profiles is $\Sigma = \{C, D\}^n$. Given a strict equilibrium $\bar{\sigma} \in \Sigma$, define

$$d^i(\bar{\sigma}, \sigma) = \left| \left\{ j \in I \mid j \neq i \text{ and } \sigma^j \neq \bar{\sigma}^j \right\} \right| \quad \text{and} \quad M^i_k(\bar{\sigma}) = \left\{ \sigma \in \Sigma \mid d^i(\bar{\sigma}, \sigma) = k \right\}$$

for $\sigma \in \Sigma$ and $k = 1, \ldots, n-1$. In words, $d^i(\bar{\sigma}, \cdot)$ counts the number of players *other than* i who play differently from the strategy in $\bar{\sigma}$. The set $M^i_k(\bar{\sigma})$ is the set of strategy profiles that contains exactly k different strategies by the others. For $\sigma = (\sigma^1, \ldots, \sigma^n) \in \Sigma$, $i \in I$, and $\eta \in \{C, D\}$, let (σ/η) be the strategy profile constructed from σ by replacing σ^i with η, with all other components fixed. Let $\tilde{\sigma}^i$ be the strategy that differs from $\bar{\sigma}^i$. For $k = 1, \ldots, n-1$, let

$$\xi^i_k(\bar{\sigma}) = \max_{\sigma_k \in M^i_k(\bar{\sigma})} u^i(\sigma_k/\tilde{\sigma}^i) - u^i(\sigma_k/\bar{\sigma}^i), \qquad (\diamond)$$

and let $\tilde{\sigma}_k$ be the maximizer of $\xi_k^i(\bar{\sigma})$.[17] Note that $\tilde{\sigma}_k \in M_k^i(\bar{\sigma})$. Set

$$\xi_0^i(\bar{\sigma}) = u^i(\bar{\sigma}) - u^i(\bar{\sigma}/\tilde{\sigma}^i).$$

Place these objects in the context of the adaptive play with mistakes. See Figure 1 and recall the discussion in the paragraphs preceding Proposition 5. Set the original state as the T-succession of a strict equilibrium $\bar{\sigma}$, the last s-portion of which is phase 1 in Figure 1. In phase 2, some players *other than* i start making *mistakes*. Strategy profiles in phase 2 are classified into sets $M_k^i(\bar{\sigma})$ according to the *number of mistakes* they contain. $\xi_k^i(\bar{\sigma})$ is the maximum payoff advantage of a *switch* from $\bar{\sigma}^i$ to $\tilde{\sigma}^i$, when player i observes exactly k mistakes by the others at one of the periods in phase 2. The maximum advantage is realized when i observes $\tilde{\sigma}_k$. The value $\xi_0^i(\bar{\sigma})$ is known as the *deviation loss*. It is strictly positive since $\bar{\sigma}$ is a strict equilibrium. Without loss of generality, phase 2 can be represented as

$$(\overbrace{\bar{\sigma}, \ldots, \bar{\sigma}}^{x_0}, \overbrace{\tilde{\sigma}_1, \ldots, \tilde{\sigma}_1}^{x_1}, \overbrace{\tilde{\sigma}_2, \ldots, \tilde{\sigma}_2}^{x_2}, \ldots, \overbrace{\tilde{\sigma}_{n-1}, \ldots, \tilde{\sigma}_{n-1}}^{x_{n-1}}).$$

Note, for example, that profile $\tilde{\sigma}_2$ contains exactly *two mistakes* made by two players other than i, and it appears exactly x_2 times in phase 2, and so on for the other $\tilde{\sigma}_k$ and x_k. We consider whether player i can play $\tilde{\sigma}^i$ as a best response to phase 2. This is possible if and only if $(x_0, x_1, \ldots, x_{n-1})$ is feasible in the following linear program:[18]

$$\begin{aligned} &\min_{x_k} \sum_{k=0}^{n-1} k x_k \\ s.t. \quad &\sum_{k=1}^{n-1} \xi_k^i(\bar{\sigma}) x_k \geq \xi_0^i(\bar{\sigma}) x_0, \quad \sum_{k=0}^{n-1} x_k = s, \quad x_k \geq 0. \end{aligned} \tag{$P^i(\bar{\sigma})$}$$

The first constraint is satisfied if and only if strategy $\tilde{\sigma}^i$ is a best response to the sample. The second constraint ensures that the set of profiles is indeed a sample in the adaptive play. The objective function counts the total number of mistakes in the sample. The objective value of the program $P^i(\bar{\sigma})$ yields the minimally possible number of mistakes in a *first exit* achieved by strategy $\tilde{\sigma}^i$ chosen by the *first exitor* i. Note that there is a distinct program to be considered for each type of the first exitor i and each $\bar{\sigma}^i \in \{C, D\}$.

[17] Although $\tilde{\sigma}_k$ depends on $i \in I$, we do not write this dependence explicitly.

[18] Strictly speaking, we should add the integer constraint to the decision variables. By implicitly assuming that the sample size s is sufficiently large, we ignore the integer constraint throughout.

Consider a group formation game with two types that satisfies (**HD**), in which each player is either type m or type M. In what follows, we write $P^m(\bar{\sigma}, C)$ to denote the linear program in which the type of the first exitor i is m and $\bar{\sigma}^i = C$, and write $V^m(\bar{\sigma}, C)$ to denote its optimal value. Adopt analogous notations for the other programs and their optimal values. There are seven programs to consider. If $\bar{\sigma}$ is a size m equilibrium, the relevant programs are $P^m(\bar{\sigma}, C)$, $P^m(\bar{\sigma}, D)$, and $P^M(\bar{\sigma}, D)$. The *exit resistance* $r(m)$ of the m-sized equilibrium is defined by

$$r(m) = \min\{V^m(\bar{\sigma}, C), V^m(\bar{\sigma}, D), V^M(\bar{\sigma}, D)\}.$$

If $\bar{\sigma}$ is an M-sized equilibrium, we need to consider $P^m(\bar{\sigma}, C)$ and $P^M(\bar{\sigma}, C)$. In addition, we consider $P^M(\bar{\sigma}, D)$ if $\bar{\sigma}$ is an (Mm, M) equilibrium and $P^m(\bar{\sigma}, D)$ if $\bar{\sigma}$ is an (Mm, m) equilibrium. Their exit resistances are defined by

$$\begin{aligned} r(Mm, M) &= \min\{V^m(\bar{\sigma}, C), V^M(\bar{\sigma}, C), V^M(\bar{\sigma}, D)\}, \\ r(Mm, m) &= \min\{V^m(\bar{\sigma}, C), V^M(\bar{\sigma}, C), V^m(\bar{\sigma}, D)\}. \end{aligned}$$

Proof of Proposition 6

In order to compute the exit resistance, we need to solve the relevant linear programs. That is to say, we need to find the mistake minimizing sample for each of the programs. In general, the mistake minimizing sample involves *mistakes by two or more players at a time*. This means that the optimal solution of the program may have nonzero x_k, $k \geq 2$. The computation of resistances in such cases may become quite tedious. However, the computation is tractable and some sharp results can be derived if the mistake minimizing sample can be found in those samples that only contain *at most one mistake at a time*. Technically, these cases are characterized by the condition that $x_2 = x_3 = \cdots = x_{n-1} = 0$ at the optimal solution. We proceed as follows. Given a strict equilibrium σ, a payoff type $\tau \in \{m, M\}$, and a current strategy $X \in \{C, D\}$, let $\xi_k^\tau(\sigma, X)$ denote $\xi_k^\tau(\sigma)$ in the linear program $P^\tau(\sigma, X)$.

- First, we solve the program with an additional constraint that $x_2 = x_3 = \cdots = x_{n-1} = 0$. This amounts to computing the explicit value of $\xi_1^\tau(\sigma, X)$.
- Second, we derive a sufficient condition, (**SM**), under which the solution found in the first step is also optimal without the additional constraint.
- Third, under the payoff assumptions (**P1**)–(**P4**), the sufficient condition (**SM**) holds.

These three steps are formalized by the next lemmata, from which Proposition 6 follows.

Lemma A1. *Consider a group formation game with two types that satisfies* (**HD**). *Let σ_m be an m-sized equilibrium and let σ_M be an M-sized equilibrium. In each program $P^\tau(\sigma, X)$, we have the following $\xi_1^\tau(\sigma, X)$ values:*

	$P^m(\sigma_m, C)$	$P^m(\sigma_m, D)$	$P^M(\sigma_m, D)$	$P^m(\sigma_M, C)$	$P^m(\sigma_M, D)$	$P^M(\sigma_M, C)$	$P^M(\sigma_M, D)$
ξ_1^τ	$D_m^m - C_m^m$	$C_{m-1}^m - D_0^m$	0	$D_M^m - C_M^m$	$C_{M-1}^m - D_0^m$	$D_M^M - C_M^M$	$C_{M-1}^M - D_0^M$

Proof. First, consider $\xi_1^m(\sigma_M, C)$ in program $P^m(\sigma_M, C)$. Let $\sigma_M = (\overbrace{C, \ldots, C, C}^{M}, D, D \ldots, D)$ be an M-sized equilibrium. Name the players 1, 2, ..., M, $M+1$, ..., n from left to right so that the rightmost cooperator is player M and the leftmost free rider is player $M+1$, with the rightmost player being n. By Proposition 2, there are at least two type-M cooperators. By (**HD**), there is also a type-m cooperator. Assume for the moment that there are at least two type-m cooperators and that both type-m and type-M free riders exist. Thus, let us assume that players 1, 2, and $M+1$ are type-m players and that M and n are type-M players. Let player 1 be the designated player who is going to be the first exitor after observing mistakes by the others. Modulo renaming of the players, we only need to consider four strategy profiles in $M_1^m(\sigma_M)$:

$$(\sigma_M/D^2) = (\overbrace{C, D, C \ldots, C, C}^{M}, D, D, \ldots, D, D), \quad (\sigma_M/D^M) = (\overbrace{C, C, C \ldots, C, D}^{M}, D, D, \ldots, D, D),$$
$$(\sigma_M/C^{M+1}) = (\overbrace{C, C, C \ldots, C, C}^{M}, C, D, \ldots, D, D), \quad (\sigma_M/C^n) = (\overbrace{C, C, C \ldots, C, C}^{M}, D, D, \ldots, D, C).$$

Neither (σ_M/D^2) nor $((\sigma_M/D^2)/D^1)$ are successful, because the original type-M cooperators are still playing C. Similarly, neither (σ_M/D^M) nor $((\sigma_M/D^M)/D^1)$ are successful, because there remains an original type-M cooperator still playing C. Both (σ_M/C^{M+1}) and $((\sigma_M/C^{M+1})/D^1)$ are successful, because there are at least M cooperators. Similarly, both (σ_M/C^n) and $((\sigma_M/C^n)/D^1)$ are successful. Strategy profiles (σ_M/C^{M+1}) and (σ_M/C^n) are payoff-equivalent to player 1, the unique best response against which is D:

$$u^m((\sigma_M/C^{M+1})/D^1) - u^m(\sigma_M/C^{M+1}) = u^m((\sigma_M/C^n)/D^1) - u^m(\sigma_M/C^n) = D_M^m - C_M^m.$$

Hence, by definition ($\diamond$), $\xi_1^m(\sigma_M, C) = D_M^m - C_M^m$.

If σ_M contains just one type-m cooperator, we omit only (σ_M/D^2) from the consideration. If all free riders are the same type, we omit either (σ_M/C^{M+1}) or (σ_M/C^n) from the consideration. The conclusion does not change in these cases. Hence $\xi_1^m(\sigma_M, C) = D_M^m - C_M^m$.

Similar but simpler arguments show the desired conclusion in other programs. Of these, programs $P^M(\sigma_M, C)$ and $P^M(\sigma_m, D)$ deserve special mention. In the former, a similar argument as above leads to $\xi_1^M(\sigma_M, C) = D_M^M - C_M^M$, thanks to the fact that in any M-sized equilibrium there are at least three type-M cooperators. If there were an M-sized equilibrium that contains exactly two type-M cooperators, then its ξ_1^M value would be $\xi_1^M = \min\{D_M^M - C_M^M, D_{M-2}^M - D_0^M\}$, and more detailed case distinctions would be required.

Finally, consider $P^M(\sigma_m, D)$. In σ_m, all cooperators are of type m. Let player $1, \ldots, m$ be cooperators. Because $|I_M| \geq 2$ and $|I_m| > m$, we can assume that player $m+1$ is a type-m free rider and player $n-1$ and n are both type-M free riders. Let n be the designated first exitor in the program. Modulo renaming of the players, we only need to consider three strategy profiles in $M_1^M(\sigma_m)$:

$$(\sigma_m/D^1) = (\overbrace{D, C \ldots, C}^{M}, D, D, \ldots, D, D, D), \quad (\sigma_m/C^{m+1}) = (\overbrace{C, C \ldots, C}^{M}, C, D, \ldots, D, D, D),$$
$$(\sigma_m/C^{n-1}) = (\overbrace{C, C \ldots, C}^{M}, D, D, \ldots, D, C, D).$$

There are no strategy profile against which C is a unique best response for player n. For (σ_m/D^1), the payoffs for strategies C and D are the same. Hence, $\xi_1^M(\sigma_m, D) = 0$. □

Lemma A2. *Consider program $P^i(\bar{\sigma})$. Write $\xi_k = \xi_k^i(\bar{\sigma})$. If*

$$\frac{\xi_k + \xi_0}{\xi_1 + \xi_0} \leq k, \tag{SM}$$

for $k \geq 2$, then the optimal value $V = V^i(\bar{\sigma})$ is given by $V = s\xi_0/(\xi_0 + \xi_1)$.

Proof. The dual of the program $P^i(\bar{\sigma})$ is the following:

$$\max s\mu \quad s.t. \quad \mu \leq \lambda\xi_0, \quad \lambda\xi_k + \mu \leq k, \ (k = 1, \ldots, n-1), \quad \lambda \geq 0,$$

where λ and μ are dual variables, the latter of which is unrestricted in sign. In program $P^i(\bar{\sigma})$, consider

$$x^* = (x_0^*, x_1^*, x_2^*, \ldots, x_{n-1}^*) = \left(\frac{s\xi_1}{\xi_0 + \xi_1}, \frac{s\xi_0}{\xi_0 + \xi_1}, 0, \ldots, 0\right).$$

The solution x^* is clearly feasible, and its objective value is $s\xi_0/(\xi_0 + \xi_1)$. In the dual program, consider

$$y^* = (\lambda^*, \mu^*) = \left(\frac{1}{\xi_0 + \xi_1}, \frac{\xi_0}{\xi_0 + \xi_1}\right),$$

whose dual objective value is equal to the preceding primal objective value. Thus, from the duality theorem, it follows that x^* is primal optimal if and only if y^* is dual feasible. That is,

$$\frac{\xi_0}{\xi_0+\xi_1} \leq \left(\frac{1}{\xi_0+\xi_1}\right)\xi_0, \quad \left(\frac{1}{\xi_0+\xi_1}\right)\xi_k + \frac{\xi_0}{\xi_0+\xi_1} \leq k, \quad (k=1,\ldots,n-1).$$

Therefore, x^* is primal optimal if and only if (**SM**) for every $k=2,\ldots,n-1$. □

Lemma A3. *Consider a group formation game with two types that satisfies* (**HD**)*. If* (**P1**)–(**P4**) *are satisfied, then the condition* (**SM**) *holds in each of the seven relevant programs.*

Proof. To see whether (**SM**) holds, it suffices to consider $k \geq 2$ such that $\xi_k^\tau(\sigma, X) > 0$. Consider program $P^m(\sigma_M, C)$. Let player m be the designated first exitor whose type is m. For some $k \geq 2$, assume that

$$\xi_k^m = \xi_k^m(\sigma_M, C) = u^m(\tilde{\sigma}_k/D^m) - u^m(\tilde{\sigma}_k/C^m) > 0.$$

Thus, D is a unique best response against $\tilde{\sigma}_k$ for player m. Hence, $(\tilde{\sigma}_k/D^m)$ must be successful. Because player m is of type m, $(\tilde{\sigma}_k/C^m)$ is also successful.

Consider the set $S = S(\tilde{\sigma}_k/D^m)$ of cooperators, where a type-M player may exist in S. Consider first the case that no type-M player exists in S. In this case,

$$\xi_k^m = D_{|S|}^m - C_{|S|}^m$$

and $m \leq |S| \leq |I_m| - 1$. Denote $|S| = M - \tilde{k}$. Because S arises with k mistakes from σ_M, $k \geq \tilde{k} - 1$. By (**HD**), $m \leq M - \tilde{k} \leq M - 4$. By Lemma A1, $\xi_1^m = \xi_1^m(\sigma_M, C) = D_M^m - C_M^m > 0$. By (**P1**) and $k \geq \tilde{k} - 1$,

$$\frac{\xi_k^m}{\xi_1^m} = \frac{D_{M-\tilde{k}}^m - C_{M-\tilde{k}}^m}{D_M^m - C_M^m} \leq |-\tilde{k}+1| = \tilde{k} - 1 \leq k.$$

Therefore, it follows from $\xi_0^m = \xi_0^m(\sigma_M, C) > 0$ that[19]

$$\frac{\xi_k^m + \xi_0^m}{\xi_1^m + \xi_0^m} \leq k.$$

Next, consider the case that S contains a type-M player. Then, $\xi_k^m = D_{|S|}^m - C_{|S|}^m$ and $M \leq |S| \leq n-1$. Denote $|S| = M + \tilde{k}$. Because S arises with k mistakes from σ_M, $k \geq \tilde{k} + 1$. By (**P1**) and $k \geq \tilde{k} + 1$,

$$\frac{\xi_k^m}{\xi_1^m} = \frac{D_{M+\tilde{k}}^m - C_{M+\tilde{k}}^m}{D_M^m - C_M^m} \leq |\tilde{k}+1| = \tilde{k} + 1 \leq k,$$

[19] Let $a > 0$, $b, c \geq 0$, and $k \geq 1$. If $b/a \leq k$, $(b+c)/(a+c) \leq k$.

which implies (**SM**). A similar but simpler argument, together with (**P1**) or (**P2**), shows the desired conclusion for $P^m(\sigma_m, C)$ or $P^M(\sigma_M, C)$, respectively.

For programs $P^m(\sigma_m, D)$, $P^m(\sigma_M, D)$, $P^M(\sigma_m, D)$, and $P^M(\sigma_M, D)$, the cases in focus are the strategy profiles against which strategy C is a unique best response to the designated first exitor. These are strategy profiles to which the first exitor is critical. If the first exitor is a type-m player, the number of cooperators in such a profile is either m or M. If the first exitor is a type-M player, the number of cooperators is M.

Consider $P^m(\sigma_m, D)$. In this program, if $\xi_k^m = \xi_k^m(\sigma_m, D) > 0$, then $\xi_k^m = C_{m-1}^m - D_0^m$ or $\xi_k^m = C_{M-1}^m - D_0^m$. The first case trivially implies the desired conclusion because $\xi_1^m = C_{m-1}^m - D_0^m$ by Lemma A1. In the second case, the number k of mistakes is at least $M-m-1$. Because $\xi_0^m = D_m^m - C_m^m$, the desired conclusion (**SM**) is

$$\frac{\xi_k^m + \xi_0^m}{\xi_1^m + \xi_0^m} = \frac{C_{M-1}^m - D_0^m + D_m^m - C_m^m}{C_{m-1}^m - D_0^m + D_m^m - C_m^m} \leq M - m - 1 \leq k,$$

which is precisely (**P3**). Consider program $P^m(\sigma_M, D)$. For this program, a similar argument leads to the desired conclusion

$$\frac{\xi_k^m + \xi_0^m}{\xi_1^m + \xi_0^m} = \frac{C_{m-1}^m - D_0^m + D_M^m - C_M^m}{C_{M-1}^m - D_0^m + D_M^m - C_M^m} \leq M - m + 1 \leq k,$$

which is trivially satisfied because $C_{M-1}^m \geq C_{m-1}^m$ by (**GF2**).

Consider $P^M(\sigma_m, D)$. In this program, if $\xi_k^M > 0$ then $\xi_k^M = C_{M-1}^M - D_0^M$ and $k \geq M - m - 1$. By Lemma A1, $\xi_1^M = 0$. Thus, the desired conclusion is

$$\frac{\xi_k^M + \xi_0^M}{\xi_0^M} = \frac{C_{M-1}^M - D_0^M + D_m^M - D_0^M}{D_m^M - D_0^M} \leq M - m - 1 \leq k,$$

which is precisely (**P4**). In program $P^M(\sigma_M, D)$, $\xi_k^M > 0$ implies that $\xi_k^M = \xi_1^M = C_{M-1}^M - D_0^M$. Thus, (**SM**) is trivially satisfied. □

Proof of Proposition 6. Consider an m-sized equilibrium σ_m. By Lemmata A1–A3, the optimal values of the relevant programs are as follows:[20]

	$P^m(\sigma_m, C)$	$P^m(\sigma_m, D)$	$P^M(\sigma_m, D)$
V	$\frac{C_{m-1}^m - D_0^m}{C_{m-1}^m - D_0^m + D_m^m - C_m^m}$	$\frac{D_m^m - C_m^m}{D_m^m - C_m^m + C_{m-1}^m - D_0^m}$	1

Hence, in terms of the value μ_m^m that was introduced in Section 4.2, the exit resistance $r(m)$ is given by $r(m) = \min\{\mu_m^m, 1 - \mu_m^m\}$. For M-sized equilibria, we have the following optimal

[20] In this proof, we divide the optimal values by s.

values:

	$P^m(\sigma_M, C)$	$P^M(\sigma_M, D)$	$P^M(\sigma_M, C)$	$P^M(\sigma_M, D)$
V	$\frac{C^m_{M-1}-D^m_0}{C^m_{M-1}-D^m_0+D^m_M-C^m_M}$	$\frac{D^m_M-C^m_M}{D^m_M-C^m_M+C^m_{M-1}-D^m_0}$	$\frac{C^M_{M-1}-D^M_0}{C^M_{M-1}-D^M_0+D^M_M-C^M_M}$	$\frac{D^M_M-C^M_M}{D^M_M-C^M_M+C^M_{M-1}-D^M_0}$

In terms of the values μ^m_M and μ^M_M that were introduced in Section 4.2,

$$r(Mm, M) = \min\left\{\mu^M_M, \mu^m_M, 1-\mu^m_M\right\}, \quad r(Mm, m) = \min\left\{\mu^M_M, \mu^m_M, 1-\mu^M_M\right\}.$$

By Proposition 5, a strict equilibrium is stochastically stable if and only if it has the largest exit resistance. One can verify that for any real number x and y,

$$\max\left\{\min\{x, y, 1-x\}, \min\{x, y, 1-y\}\right\} = \min\left\{\min\{x, y\}, 1-\min\{x, y\}\right\}.$$

Thus, setting $\mu^{\tau^*}_M = \min\{\mu^m_M, \mu^M_M\}$,

$$\max\{r(Mm, M), r(Mm, m)\} = \min\{\mu^{\tau^*}_M, 1-\mu^{\tau^*}_M\}.$$

Recalling $\Pi_m = \mu^m_m(1-\mu^m_m)$ and $\Pi_M = \mu^{\tau^*}_M(1-\mu^{\tau^*}_M)$, the result follows from the fact that $p(1-p) \geq q(1-q)$ if and only if $\min\{p, 1-p\} \geq \min\{q, 1-q\}$ for real numbers p and q between zero and one. □