

Discussion Paper No. 006

THE LONG-RUN STABILITY OF GROUP FORMATION
IN COLLECTIVE ACTION GAMES

Toshimasa Maruta
and
Akira Okada

December, 2003

21COE
Interfaces for Advanced Economic Analysis
Kyoto University

THE LONG-RUN STABILITY OF GROUP FORMATION IN COLLECTIVE ACTION GAMES

Toshimasa Maruta¹ and Akira Okada²

December 12, 2003

ABSTRACT: We present a simple model of voluntary groups in collective action and investigate the long-run stability of group formation by applying the stochastic evolutionary theory. The heterogeneity of individuals' preferences yields the multiplicity of strict Nash equilibria with different levels of cooperation. The non-cooperation equilibrium is not stochastically stable when players sample a large number of past plays. When the number of players less motivated to cooperate is larger than a critical level, the partial cooperation equilibrium is uniquely stochastically stable. Otherwise, the stochastic stable equilibrium is determined according to a version of risk-dominance, under which the full cooperation can be stochastically stable.

Journal of Economic Literature Classification Numbers: C70, C72.

KEYWORDS: Collective action, equilibrium selection, evolution, group formation, heterogeneous preferences, prisoner's dilemma, stochastic stability.

¹Faculty of Economics, Tokyo Metropolitan University, 1-1 Minami Osawa, Hachioji, Tokyo 192-0397 Japan.
E-mail: tosh@comp.metro-u.ac.jp Phone: +81 426 77 1138 Fax: +81 426 77 2304

²Corresponding author: Institute of Economic Research, Kyoto University, Sakyo, Kyoto 606-8501 Japan.
E-mail: okada@kier.kyoto-u.ac.jp Phone: +81 75 753 7145 Fax: +81 75 753 7148

1 Introduction

The collective action problem arises in social situations in which it is a common interest for all individuals to act collectively but each of them has an incentive to free-ride on the others' actions. Examples of collective action problems can be found in many social, political, and economic activities. They include: public goods provision, cartel formation, labor union, environmental pollution, common-pool resources management, participation in community work, trading associations, international organizations, etc. The central question in the collective action problem is how a group of individuals voluntarily cooperate, and how the voluntary group can be sustained in the long-run. The purpose of this paper is to present a game theoretic model of voluntary groups in collective action, and to investigate the long-run stability of group formation by applying the stochastic evolutionary theory introduced by Young (1993) and Kandori, Mailath and Rob (1993). Modelling the collective action problem as an n -person prisoner's dilemma, our analysis focuses on how the heterogeneity of individuals' preferences affects the formation and the stability of voluntary groups in collective action.

In many real situations, individuals differ in their willingness to participate in collective action. For example, consider a problem of environmental pollution. Some individuals are concerned very much with environmental preservation, and they are willing to contribute for the prevention of pollution even if they have a small number of followers. Others may be reluctant to participate in such a collective activity. They might contribute for the activity only if a large number of individuals have already done or will do. In such heterogeneous situations, we ask the following questions: Does a voluntary group consist only of individuals with higher willingness to cooperate, or does it include many types of individuals? If different kinds of groups are possible, which one is stable in the long-run?

The process of group formation is modeled as a two-stage game. In the first stage, individuals decide independently to participate in a group or not. In the second stage, participants negotiate on the collective action. The agreement of cooperation is reached only if all participants agree to do so. If the agreement is reached, then participants are bound to cooperate, bearing the group costs. Any non-participant is allowed to free-ride. If the agreement is not reached, then the group breaks up, and the n -person prisoner's dilemma is played.

Each individual's incentive to cooperate is characterized by the minimum size of a group in which her participation to it can make her better off (even with bearing group costs) than in

the non-cooperative equilibrium of the prisoner’s dilemma. In this paper, we call the minimum size the individual’s *threshold of cooperation*. Individuals with smaller thresholds are more motivated to cooperate. It is shown that a group is formed in a Nash equilibrium of the second stage game if and only if the group size exceeds thresholds of its members. Such a group is called *successful*. By solving backward, the two-stage group formation game is reduced to the following n -person strategic-form game, which we call the *group formation game*. In this game, all individuals decide independently to participate in a group or not. The group is formed if and only if it is successful, that is, all participants are better off by cooperating than in the non-cooperative equilibrium. Any non-member free-rides on the group action.

In the first part of the paper, we show that the group formation game with heterogeneous preferences has multiple strict Nash equilibria. Specifically, besides the non-participation equilibrium, there are multiple equilibria that differ from each other in group size. Any successful group can be formed in equilibrium whenever it is also “critical,” in that any unilateral deviation makes it unsuccessful. With heterogeneous preferences, there are typically many such groups, ranging from a group with just a few members, to the group formed by all individuals. If preference were homogeneous, in contrast, equilibrium group size would be uniquely determined. In this sense, the multiplicity of equilibrium group sizes is a salient feature of heterogeneous preferences. Having such many equilibrium groups, one naturally asks which group is likely to be formed.

It should be remarked that a multiplicity of similar kind arises in the repeated game analysis of the n -person prisoner’s dilemma. Specifically, any cooperative group in which each member receives an individually rational payoff can be supported as a subgame perfect equilibrium.¹ In general, there are many such equilibrium groups. On the play path in such an equilibrium, every non-member free-rides. Although one could ask which equilibrium is likely to prevail, the issue has remained relatively unexplored in the repeated game literature.

In the second part of the paper, we consider the equilibrium selection problem of the group formation game. To investigate which equilibrium can be sustained in the long-run, we apply the adaptive play model (Young 1993) to the group formation game. In the adaptive play, each individual plays a best response to a sample, which is a set of action profiles, randomly selected

¹For example, consider a trigger strategy as follows (Taylor 1987). Every member cooperates first, and continues to cooperate as long as everyone else in the group continues to do so, but defects forever after once any deviation occurs. Meanwhile, all non-members always defect.

from recent plays. From time to time, however, a player may fail to choose a best response, but may end up with a random strategy choice with a small probability. It turns out that the size of the sample is one of the crucial parameters in the subsequent analysis.

When we observe the adaptive play long enough, some states distinguish themselves from others in that they are observed almost all the time. Such states are called stochastically stable, and we examine the stochastic stability of Nash equilibria. To make the result transparent, we restrict the analysis to a special case in which there are exactly two types of individuals. The first types have a common threshold of cooperation, which is well below the whole population size. The second types, in contrast, have the largest possible thresholds, namely the population size. The first types might cooperate even if the second types do not. The second types are the least motivated to cooperate, in that each of them has an incentive to cooperate only if all the others do. In the two-type model, there are three strict Nash equilibria: In non-cooperation equilibrium, none cooperates; In partial cooperation equilibrium, only the first types cooperate; In full cooperation equilibrium, both types cooperate. In what follows, we call the second type of individuals *potential free-riders*, because it is the second types that might free-ride in equilibrium. The two-type model helps us obtain a basic insight about how different levels of cooperation can be sustained in the long-run.

The main results of the paper are as follows. First, the selection outcome depends on the number of first types and the sample size in the adaptive play. When the former exceeds the latter, the non-cooperation equilibrium is stochastically stable (Theorem 4.1). When it does not, then the outcome depends on the number of second types, that is, the number of potential free-riders, and the strength of their incentives to cooperate. If the number of potential free-riders exceeds a critical level, then the partial cooperation equilibrium is stochastically stable (Theorem 4.2). Finally, if the number of potential free-riders is below the critical level, a version of risk-dominance (Harsanyi and Selten 1988) determines the selection outcome (Theorem 4.3). The analysis in this case can be reduced to the 2×2 coordination game played by the two potential free-riders, who have the highest and the lowest incentive to cooperate, respectively. As long as both the highest and the lowest levels of incentive remain unchanged, the distribution of intermediate levels of incentive is irrelevant to the equilibrium selection.

To focus our analysis on the voluntary formation of groups, we leave the enforcement problem of cooperation in a group out of the scope of this paper. We simply assume that the group,

once formed, is endowed with some costly enforcement mechanism. Enforcement mechanisms of cooperation have been extensively studied in the literature. The well-studied mechanism is the mutual punishments among participants in repeated games. Under suitable conditions, the folk theorem of repeated games (e.g., Fudenberg and Maskin 1986) shows that a collective action can be enforced by conditional behavioral rules, such as trigger strategies. Another type of mechanism is a centralized institution created by participants themselves. The institutional arrangement approach assumes that participants use scarce resources to create an enforcement institution and transfer the right of punishing defectors to enforcers, who may or may not be chosen from the participants. We remark that whatever enforcement mechanism is employed, the mechanism itself is a kind of public good and that every individual has an incentive to free-ride on it. This problem has been called the second order dilemma in the provision of public goods (Ostrom 1998). There exists a strategic conflict among individuals regarding who participate in the mechanism of cooperation. Our group formation game is a simple model to analyze the voluntary participation in an enforcement mechanism of cooperation.

This paper is related to several works in the literature. Since the classic work of Olson (1965), the group formation in collective action has been widely investigated. The group size effect, argued by Olson (1965), that larger groups are less successful in organizing collective action is not necessary true in our model. The success of collective action critically depends upon the benefit and the cost to participate. The largest group can be formed in a unique long-run equilibrium. Our model differs from other “threshold” or “critical mass” models (Schelling 1978, Granovetter 1978, Marwell and Oliver 1988) in formulating individuals’ behavior. Such models typically presume a simple behavioral rule of individuals, in which they are programmed to participate in collective action whenever the number of participants exceeds their thresholds. On the other hand, individuals behave strategically in our group formation game. That is, even if the current number of participants exceeds their thresholds, they do not automatically participate, since their best responses in such a situation might be to free-ride. Adopting Palfrey and Rosenthal’s (1984) model of public goods, Diermeier and van Mieghem (2000) study a stochastic process of collective action. Working in a homogeneous population with a log-linear choice rule, they characterize the limit distribution of their birth and death process model. Finally, our work is also related to recent studies on network formation. Goyal and Vega-Redondo (2000) present a noncooperative model of network formation in which individuals

choose partners and a linked pair of individuals play 2×2 coordination games. They investigate the stochastic stability of network formation and of coordination, and show that costs of forming links critically affect the equilibrium selection results. In their model, link formation is one-sided such that a link can be made by the initiative of one partner, who bears its costs, and it is always accepted by the other partner. In contrast, we analyze the strategic conflict of group formation, in which the group formation involves a coordination problem.

The paper is organized as follows. Section 2 presents a group formation game. Section 3 characterizes strict Nash equilibria of the game. Section 4 investigates the equilibrium selection problem in the group formation game. Results in this section are proven in the Appendix. We conclude in Section 5.

2 The Model

Consider an n -person prisoner's dilemma. Let $N = \{1, 2, \dots, n\}$ be the set of players. Every player $i \in N$ has two actions, C (cooperation) and D (defection). Player i 's payoff is given by

$$u_i(a_i, h), \quad a_i = C, D, \quad h = 0, 1, \dots, n-1,$$

where a_i is player i 's action and h is the number of other players who select C . We make the following assumption.

Assumption 2.1. The payoff function of player $i \in N$ satisfies:

- (1) $u_i(D, h) > u_i(C, h)$ for every $h = 0, 1, \dots, n-1$,
- (2) $u_i(C, n-1) > u_i(D, 0)$,
- (3) $u_i(C, h)$ and $u_i(D, h)$ are increasing in h .

This assumption is standard in the literature of an n -person prisoner's dilemma (Schelling 1978), except that players are "heterogeneous" in the sense that they have different payoff functions. The heterogeneity of players is critical to the analysis of this paper. Property (1) means that every player is better off by defecting than cooperating, regardless of others' play. This implies that every player has an incentive to free-ride on others' cooperative action. Thus, the action profile (D, \dots, D) is a unique Nash equilibrium of the game. On the other hand, property (2) says that if all players cooperate, they are all better off than in the unique

equilibrium. Namely, the equilibrium is not Pareto efficient. Property (3) means that the more others cooperate, the higher payoff each player can receive, regardless of her action. The cooperative action by any player has positive externality on all others.

The prisoner's dilemma describes an anarchic situation in which players are free to choose their actions. In such a situation, a natural outcome of the game is the Nash equilibrium in which no one cooperates. The central question concerning the prisoner's dilemma is whether and how self-interested individuals voluntarily cooperate under the presence of temptations to defect. To escape from an undesirable state of non-cooperation, some suitable mechanisms for preventing opportunistic behavior are needed. The literature has considered diverse mechanisms attaining cooperation. However, as we have discussed in the introduction, any mechanism of cooperation itself is a kind of public good. Every individual has an incentive to free-ride on the mechanism. We formulate the following game of group formation to consider whether or not individuals voluntarily participate in the mechanism of cooperation.

The process of group formation is defined as a two-stage game.

Participation decision stage: Every player $i \in N$ decides independently whether or not to participate in a group, to negotiate on the collective action. Participation takes some costs, say, for phone calls, mails and transportations. The participation cost is denoted by a small positive amount $\varepsilon_i > 0$. Let S be the set of all s participants. If $s = 0$ or $s = 1$, then no group is possible.²

Group negotiation stage: All participants negotiate on their cooperation according to the unanimity rule. Each of them either accepts or rejects independently to cooperate. The agreement of cooperation is reached if and only if all participants accept it. When the agreement of cooperation is reached, it is enforced and all participants are bound to cooperate. The enforcement is costly and all participants have to bear the group costs (including participation costs ε_i). The group cost per capita is given by a real valued function $c(s)$ where s is the number of all participants. All non-participants are free to defect. When an agreement is not reached, all individuals, both participants and non-participants, play the original prisoner's dilemma.

When all participants agree to cooperate, we assume that the group is endowed with some mechanism to enforce the agreement. The mechanism has various functions such as monitoring

²When $s = 1$, the single participant has no incentive to cooperate in the prisoner's dilemma.

members' actions and punishing defecting members. Obviously, it is costly for group members to have such a mechanism. In what follows, to keep our game model as simple as possible, we do not present a formal model of an enforcement mechanism in a group. Rather, we formulate it simply by a group cost function $c(s)$. Okada (1993) considers a related model of collective action in which one participant is selected as an enforcer.

We now consider a subgame perfect equilibrium of the two-stage game of group formation. By backward induction, we first analyze the group negotiation stage. When a group of s members agree to cooperate, every member receives utility

$$v_i(C, s - 1) \equiv u_i(C, s - 1) - c(s).$$

We call $v_i(C, s - 1)$ the *group payoff* of player i where s is the number of group members. Concerning the group payoff, we assume the following property.

Assumption 2.2. For every $i \in N$, the group payoff $v_i(C, s - 1)$ of player i is monotonically increasing in s , and there exists a unique integer s_i ($2 \leq s_i \leq n$) such that

$$v_i(C, s_i - 2) < u_i(D, 0) < v_i(C, s_i - 1). \quad (2.1)$$

This assumption means that even if we replace the original cooperative payoff $u_i(C, h)$ with the group payoff $v_i(C, h)$, the properties (Assumption 2.1) of the n -person prisoner's dilemma still hold true. If Assumption 2.2 does not hold, the problem of group formation becomes trivial. For example, if $v_i(C, s - 1) \leq u_i(D, 0)$ for all $s \leq n$, then no players have incentive to participate in a group. The positive integer s_i in (2.1) shows the minimum size of a group in which member i can be better off than in the non-cooperative equilibrium of the prisoner's dilemma. We call s_i the *threshold of cooperation* of i . Player i can benefit by cooperating whenever at least $(s_i - 1)$ others cooperate. In this sense, players with smaller thresholds have higher motivation to cooperate.

Definition 2.1. A subset S of N is called a *successful group* if $|S| \geq s_i$ for every $i \in S$.

The size of a successful group is greater than or equal to all members' thresholds of cooperation. By definition, every member of a successful group can receive higher payoff than the non-cooperative payoff in the prisoner's dilemma. The naming of a successful group is explained by the following proposition.

Proposition 2.1. *In the group negotiation stage, an agreement of cooperation is reached in a strict Nash equilibrium if and only if the group of participants is successful.³*

Proof. Suppose that all s participants agree to cooperate. Then, every participant receives the group payoff $v_i(C, s - 1)$. If any member rejects to cooperate, the negotiation breaks down by the unanimity rule, and she receives the non-cooperative payoff $u_i(D, 0)$. Therefore, the agreement of cooperation in a group S is reached in a strict Nash equilibrium if and only if for all $i \in S$, $v_i(C, |S| - 1) > u_i(D, 0)$. From Assumptions 2.1 and 2.2, this is equivalent to that the group is successful. \square

Given the Nash equilibrium of the group negotiation stage, the whole two-stage game can be reduced to the following game. In this game, every player $i \in N$ chooses simultaneously and independently either $\sigma_i = 1$ (participation) or $\sigma_i = 0$ (non-participation). Let $\Sigma_i = \{0, 1\}$ be the set of actions of player i , and let $\Sigma = \prod_{i \in N} \Sigma_i$ be the set of action profiles of n players. For an action profile $\sigma = (\sigma_1, \dots, \sigma_n) \in \Sigma$, the set $S(\sigma)$ of participants is given by

$$S(\sigma) = \{i \in N | \sigma_i = 1\}.$$

The payoff $f_i(\sigma)$ of player i for an action profile $\sigma = (\sigma_1, \dots, \sigma_n) \in \Sigma$ is defined as follows.

(i) When a group $S(\sigma)$ of participants is successful,

$$f_i(\sigma) = \begin{cases} v_i(C, |S(\sigma)| - 1) & \text{if } \sigma_i = 1, \\ u_i(D, |S(\sigma)|) & \text{if } \sigma_i = 0. \end{cases}$$

(ii) When $S(\sigma)$ is not successful,

$$f_i(\sigma) = \begin{cases} u_i(D, 0) - \varepsilon_i & \text{if } \sigma_i = 1, \\ u_i(D, 0) & \text{if } \sigma_i = 0, \end{cases}$$

where $\varepsilon_i > 0$ is a participation cost.

³There exist many “trivial” non-strict Nash equilibria leading to the disagreement. For example, any action profile where at least two participants reject to cooperate is such an equilibrium. These equilibria are peculiar to the unanimity rule where everyone has a veto power. We remark that from the viewpoint of each participant the action of agreement (weakly) dominates that of disagreement. For this reason, we only consider strict Nash equilibria leading to the agreement of cooperation in every successful group.

Formally, the process of group formation reduces to the n -person game $\Gamma = (N, \{\Sigma_i, f_i\}_{i \in N})$ in strategic form. We call it the *group formation game*. The group formation game Γ differs from the original n -person prisoner's dilemma. In Γ , non-participation does not dominate participation, nor vice versa. In particular, a non-empty group of participants may arise in equilibrium, in which non-participants free-ride on the group action. Every participant, however, can guarantee the equilibrium payoff in the prisoner's dilemma.

3 The Nash Equilibria in the Group Formation Game

In this section, we characterize the Nash equilibria of the group formation game Γ . To do this, we examine the best response structure of Γ . For an action profile $\sigma = (\sigma_1, \dots, \sigma_n) \in \Sigma$, let σ_{-i} be the action profile obtained from σ by deleting σ_i . As usual, an action profile $\sigma = (\sigma_1, \dots, \sigma_n)$ is sometimes denoted by $\sigma = (\sigma_{-i}, \sigma_i)$. Let $S(\sigma)$ be the set of participants in σ .

Definition 3.1. For an action profile $\sigma = (\sigma_{-i}, \sigma_i) \in \Sigma$ in Γ , player i 's action σ_i is called a *best response* to σ if $f_i(\sigma_{-i}, \sigma_i) = \max_{\sigma'_i \in \Sigma_i} f_i(\sigma_{-i}, \sigma'_i)$.

Definition 3.2. The *best response graph* V of Γ is a binary relation on the set of action profiles Σ such that, for every $\sigma, \sigma' \in \Sigma$, $(\sigma, \sigma') \in V$ if and only if $\sigma \neq \sigma'$ and there exists exactly one player i satisfying (i) $\sigma_{-i} = \sigma'_{-i}$ and (ii) σ'_i is a best response to σ for i . When $(\sigma, \sigma') \in V$, we write $\sigma \rightarrow \sigma'$ and call it an *edge* from σ to σ' .

The definition of the best response graph is due to Young (1993).

Definition 3.3. For a successful group S , member i of S is called *critical* if $S \setminus \{i\}$ is not successful, and *non-critical* if $S \setminus \{i\}$ is successful.

No successful group can sustain itself if any critical member opts out of it. The following proposition characterizes the best response graph of the group formation game Γ .

Proposition 3.1. *An edge of the best response graph V of the group formation game Γ must be one of the following types. Let $\sigma \in \Sigma$.*

- (1) *Assume that $S(\sigma)$ is successful. For every $\sigma' \in \Sigma$, $\sigma \rightarrow \sigma'$ if and only if there is a non-critical $i \in S(\sigma)$ such that $\sigma = (\sigma_{-i}, 1)$ and $\sigma' = (\sigma_{-i}, 0)$.*

- (2) Assume that $S(\sigma)$ is not successful. For every $\sigma' \in \Sigma$, $\sigma \rightarrow \sigma'$ if and only if either $\sigma = (\sigma_{-i}, 1)$ and $\sigma' = (\sigma_{-i}, 0)$ for some $i \in S(\sigma)$ or $\sigma = (\sigma_{-i}, 0)$ and $\sigma' = (\sigma_{-i}, 1)$ for $i \notin S(\sigma)$ such that $S(\sigma) \cup \{i\}$ is successful.

Proof. (1) Suppose that $S(\sigma)$ is a successful group and that member i is not critical. Since the group $S(\sigma) \setminus \{i\}$ remains successful, we have

$$f_i(\sigma_{-i}, 1) = v_i(C, |S(\sigma)| - 1) < u_i(D, |S(\sigma)| - 1) = f_i(\sigma_{-i}, 0).$$

Therefore, $\sigma_i = 0$ is a best response to σ for all non-critical members i of $S(\sigma)$.

- (2) Suppose that $S(\sigma)$ is not a successful group. Then, for all $i \in S$,

$$f_i(\sigma_{-i}, 1) = u_i(D, 0) - \varepsilon_i < u_i(D, 0) \leq f_i(\sigma_{-i}, 0),$$

where $f_i(\sigma_{-i}, 0)$ is equal to either $u_i(D, |S(\sigma)| - 1)$ or $u_i(D, 0)$, depending on whether the remaining group except player i is successful or not. For any non-member i such that $S(\sigma) \cup \{i\}$ is a successful group,

$$f_i(\sigma_{-i}, 0) = u_i(D, 0) < v_i(C, |S(\sigma)|) = f_i(\sigma_{-i}, 1).$$

Finally, it can be easily seen that there exist no other edges in the best response graph V given in the theorem. \square

Proposition 3.1 reveals the best response structure of the group formation game. In a successful group, every non-critical member has an incentive to deviate from the group because, by doing so, she can free-ride on the group. In an unsuccessful group, every member has an incentive to deviate from the group for saving participation costs. Remark that a player outside the unsuccessful group has an incentive to join the group if her participation makes the group successful. By Proposition 3.1, we can characterize strict Nash equilibria in Γ .

Proposition 3.2. *The group formation game Γ has the following strict Nash equilibria $\sigma = (\sigma_1, \dots, \sigma_n)$.*

- (1) $\sigma = (0, \dots, 0)$, i.e., $S(\sigma) = \emptyset$.
- (2) $S(\sigma)$ is a successful group with every member of it being critical.

Since the proof is straightforward, we omit it. The proposition can be explained intuitively by an alternative definition of a Nash equilibrium in the group formation game. A group of participants in a Nash equilibrium satisfies two stability properties:

Internal stability: No single member wants to opt out of the group.

External stability: No single outsider wants to join the group.

It is clear that the action profile $\sigma = (0, \dots, 0)$ is a Nash equilibrium because no one is willing to cooperate unilaterally. When a group is not successful, the internal stability is violated because all participants want to opt out of the group for saving participation costs. When a group is successful, the external stability always holds because all non-participants have incentive to free-ride. The internal stability implies that every participant is critical. Note that if there exists no participation costs, the action profile $\sigma = (0, \dots, 0)$ is a non-strict Nash equilibrium because both actions give her the same payoff.

Another consequence of Proposition 3.1 is that the group formation game Γ is *weakly acyclic* in the sense of Young (1993). This result provides a basis for the analysis in Section 4.

Proposition 3.3. *For every action profile σ^1 in the group formation game Γ , there is a sequence of action profiles $\sigma^1 \rightarrow \dots \rightarrow \sigma^L$ in the best response graph V such that σ^L is a strict Nash equilibrium.*

Proof. Take σ^1 and assume first that it is successful. By Proposition 3.2, we can assume that there are non-critical members in $S(\sigma^1)$. Proposition 3.1 implies that $\sigma^1 \rightarrow \sigma^2$ iff $\sigma^2 = (\sigma_{-i}^1, 0)$, where $i \in S(\sigma^1)$ is not critical. Since i is not critical, Definition 3.1 implies that $S(\sigma^2)$ is successful. By induction, in any sequence $\sigma^1 \rightarrow \dots \rightarrow \sigma^L$ in V , $S(\sigma^l)$ is successful for every $l = 1, \dots, L$ and the number of players who choose action 1 is strictly decreasing. Thus the length of such a sequence is at most $|S(\sigma^1)| + 1$. Take the longest such sequence $\sigma^1 \rightarrow \dots \rightarrow \sigma^{L^*}$. Then σ^{L^*} is successful and every member of $S(\sigma^{L^*})$ is critical. That is, σ^{L^*} is a strict Nash equilibrium.

Assume next that σ^1 is not successful and take a sequence $\sigma^1 \rightarrow \dots \rightarrow \sigma^L$ in V . If there is l ($l = 2, \dots, L$) such that σ^l is successful, then the above paragraph applies. Thus we can assume that every σ^l is not successful. By Proposition 3.1 $\sigma^{l+1} = (\sigma_{-i}^l, 0)$ for $l = 1, \dots, L - 1$. Thus the number of 1-players is strictly decreasing, and the length of such a sequence is at

most $L^* = |S(\sigma^1)| + 1$. Take the longest such sequence $\sigma^1 \rightarrow \dots \rightarrow \sigma^{L^*}$. It is clear that σ^{L^*} is the non-cooperation equilibrium. \square

We can derive yet another characterization of Nash equilibria in the group formation game. It is given in terms of players' thresholds of cooperation. For $S \subset N$ and $m = 2, \dots, n$, we define $F_S(m)$ by the number of all members in S whose thresholds of cooperation are given by m . That is, $F_S(m) = |\{i \in S \mid s_i = m\}|$. F_S represents the distribution of members' thresholds of cooperation in the group S . Its definition implies that:

Lemma 3.1. *For $S \subset N$,*

$$(1) \quad F_S(2) + \dots + F_S(|S|) \leq |S|.$$

$$(2) \quad \text{A group } S \text{ is successful if and only if } F_S(2) + \dots + F_S(|S|) = |S|.$$

Proposition 3.4. *A nonempty subset S of N is the set of participants in a Nash equilibrium of the group formation game Γ if and only if*

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

Proof. From Proposition 3.2 and Lemma 3.1, it suffices to show that every member of 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. If $F_S(|S|) = 1$, then a unique member i with $s_i = |S|$ is not critical because $S \setminus \{i\}$ is a successful group. If $F_S(|S|) = 0$, all members j of S have thresholds s_j of cooperation with $s_j \leq |S| - 1$. Therefore, they are not critical. \square

The proposition allows us to see how the heterogeneity of individuals affects the group formation in collective action. When individuals are so homogeneous that they have identical thresholds, the size of an equilibrium group is uniquely determined by the common threshold. In contrast, when individuals are heterogeneous, there typically exist many equilibrium groups with different sizes. For example, if the distribution F_N of all individuals' thresholds of cooperation has a wide support so that for each integer $s = 2, \dots, m$ there exist at least two individuals with thresholds s , then groups of all sizes $s = 2, \dots, m$ can be formed in equilibrium. The heterogeneity of individuals causes the multiplicity of equilibrium groups.

The heterogeneous thresholds also affect the likelihood of the universal cooperation. In the homogeneous case, the group of n players can be formed under a stringent condition that all individuals' thresholds of cooperation are equal to the number n of players. To put it differently, the full cooperation can be attained among homogeneous individuals only if they happen to have the largest possible thresholds, namely $F_N(n) = n$. In contrast, the largest group can be sustained in equilibrium among heterogeneous individuals under a much weaker condition that $F_N(n) \geq 2$.

4 Equilibrium Selection in Group Formation Games

The analysis thus far shows that there are multiple equilibria in a group formation game with heterogeneous preferences. More specifically, there are three types of equilibria. First, the non-cooperation is always a strict equilibrium, in which no one cooperates. Second, there are typically partial cooperation equilibria, in which some players cooperate but the others do not. Third, the full cooperation equilibrium in which everyone cooperates. Thus the question arises as to which type of equilibrium is most likely to prevail in the long-run.

To tackle this problem, we adopt the stochastic equilibrium selection theory à la Young (1993). Given a strategic form game, consider an adaptive strategy revision process in which each individual plays a best response to a *sample* of past plays, where a sample is a set of k action profiles, each of which is randomly drawn without replacement from m recent action profiles. From time to time, however, a player may fail to choose a best response, but may end up with a random strategy choice with probability $\epsilon > 0$. If the randomly chosen action is not a best response to any sample that might be drawn, then the action is called a *mistake*. Such a process is called an *adaptive play with mistakes* by Young (1993). It turns out that the size k of the sample is one of the crucial parameters in the subsequent analysis.

Since the model gives rise to a Markov chain on the set of action profiles,⁴ its long term behavior can be captured by its stationary distribution μ_ϵ . In particular, we are interested in the limiting distribution $\mu = \lim_{\epsilon \rightarrow 0} \mu_\epsilon$ as the probability of mistakes vanishes. States to which the limiting distribution μ puts positive probability are called *stochastically stable*, and

⁴In Young's (1993) analysis, a *state* is formally defined to be a sequence of action profiles with length m . Thus the state space of the Markov chain is the m -fold product of the set of action profiles.

a strict equilibrium of the underlying game is stochastically stable if the corresponding state⁵ is stochastically stable. Young (1993) shows that stochastically stable states correspond to strict equilibria when the stage game is weakly acyclic in the sense of Proposition 3.3. Thus a unique stochastically stable state in the adaptive play of the group formation game corresponds to the equilibrium that is observed infinitely more times than others in the long-run, when the probability of mistakes is vanishingly small. For details, the reader is referred to Young (1993).

Applying the selection theory of Young (1993), we can identify, in principle, which equilibrium is the most stable in the sense of stochastic stability. In practice, however, it turns out to be quite complex to do so for a general group formation game. Thus we restrict our attention to *group formation games with two types of players*.

4.1 The Two Types Model

A group formation game with two types of players is defined as follows. The set $N = \{1, \dots, n\}$ of players is partitioned into $N_1 = \{1, \dots, n_1\}$ and $N_2 = \{n_1 + 1, \dots, n\}$. The size of N_1 and N_2 are n_1 and $n_2 = n - n_1$, respectively. We assume that $n_1, n_2 \geq 2$. N_1 and N_2 represent two types of players as follows. Assume that

$$s_i = n_1 \text{ for every } i \in N_1, \text{ and } s_i = n = n_1 + n_2 \text{ for every } i \in N_2,$$

where s_i is player i 's threshold of cooperation, defined in Section 2. In the notation in Proposition 3.4,

$$F_S(m) = \begin{cases} |S \cap N_1|, & \text{if } m = n_1, \\ |S \cap N_2|, & \text{if } m = n_1 + n_2, \\ 0, & \text{otherwise} \end{cases}$$

for every nonempty $S \subset N$. For a player in N_1 , it is optimal to cooperate when just $n_1 - 1$ others do. In contrast, a player in N_2 has lower motivation to cooperate, in that only when all the other $n_1 + n_2 - 1$ players cooperate, it becomes advantageous to herself to follow suit.

Proposition 4.1. *In a group formation game with two types, there are exactly three strict Nash equilibria. They are*

$$e^1 = (0, \dots, 0), \quad e^2 = (\overbrace{1, \dots, 1}^{n_1}, \overbrace{0, \dots, 0}^{n_2}), \quad \text{and} \quad e^3 = (1, \dots, 1).$$

⁵The corresponding state is a sequence (e, \dots, e) of length m , where e is a strict equilibrium of the underlying game.

The proposition follows from Proposition 3.4. Although the assumption of two types is certainly restricted, it allows us to significantly simplify the analysis, without losing the essential structure of strict Nash equilibria in general group formation games. In words, there are three equilibria with different levels of cooperation, the non-cooperation equilibrium e^1 , the partial cooperation equilibrium e^2 , and the full cooperation equilibrium e^3 . There is a crucial difference between different types of players. For any player in N_1 , there is no outcome in which she free-rides. In contrast, there is at least one outcome in which members of N_2 free-ride. In fact, they free-ride in the partial cooperation equilibrium e^2 . In this sense, members of N_2 are potential free-riders. As a result, their best response structure is more intricate than that of players in N_1 . For this reason, it is behavior and payoff of potential free-riders that become most critical to the subsequent analysis.

It proves useful to parameterize players' payoffs. For every $i \in N_1$, define

$$a_i = v_i(C, n_1 - 1), \quad c_i = u_i(D, 0) - \varepsilon_i, \quad d_i = u_i(D, 0).$$

a_i is the equilibrium payoff in e^2 , the partial cooperation. A unilateral deviation by an $i \in N_1$ from e^2 results in d_i , which is equal to the equilibrium payoff in e^1 , the non-cooperation. A unilateral deviation by an $i \in N_1$ from e^1 results in c_i , since she has to pay participation cost ε_i . Note that $a_i > d_i > c_i$. For every $i \in N_2$, let

$$a_i = v_i(C, n - 1), \quad c_i = u_i(D, 0) - \varepsilon_i, \quad d_i = u_i(D, 0), \quad f_i = u_i(D, n_1).$$

For player $i \in N_2$, a_i is the equilibrium payoff in e^3 , the full cooperation. A unilateral deviation by an $i \in N_2$ from e^3 results in d_i , which is equal to the equilibrium payoff in e^1 . A unilateral deviation by an $i \in N_2$ from e^1 results in c_i . f_i is the free riding payoff. Note that $f_i > d_i > c_i$ and $a_i > d_i$.

4.2 Evaluating Resistances

In an adaptive play, consider a state in which all of the past m plays are the same action profile e , which is a strict Nash equilibrium of the underlying game.⁶ If there were no mistakes, that is, $\epsilon = 0$, then the adaptive play would never leave there once it reached there. This is because, first, any sample in such a state must be a k -fold repetition of e , and second, the

⁶This is the state that corresponds to equilibrium e . See the preceding footnote.

best response to such a sample must be e itself since it is a strict equilibrium. With mistakes, however, the adaptive play can escape from the equilibrium state, thanks to an accumulation of a number of successive or simultaneous mistakes. In other words, a path, or a sequence of action profiles, that connects one equilibrium to another necessarily contains mistakes.⁷ Intuitively, the probability of a path is a decreasing function of the number of mistakes it contains. In order to identify the stochastically stable equilibrium, therefore, it is crucial to evaluate the number of mistakes on the path. In particular, the analysis hinges on the *resistance* $r(e, e')$, the minimum number of mistakes required for any path from an equilibrium e to another equilibrium e' .

Roughly speaking, the resistance from one equilibrium to another is inversely related to the likelihood for the noisy adaptive process to make that transition. The smaller the resistance, the more likely the path to be realized. To identify stochastically stable equilibrium, Young's (1993) selection theory compares resistances between equilibria. If, as in a 2×2 coordination game, there are only two equilibria, one equilibrium is stochastically stable if and only if the resistance to it is smaller than the resistance from it. For games with more than two equilibria, we need to consider directed graphs, or *trees*, on the set of equilibrium states. Each edge in a tree is weighted by the associated resistance. An equilibrium is stochastically stable if it is easiest to reach or most likely to realize, in that it is the root of the minimum resistance tree. For a group formation game, the relevant trees are depicted in Figure 1.

(Figure 1 appears about here.)

Evaluating resistances is a tedious exercise. Detailed arguments on resistance are delegated to the Appendix. In this section, we exhibit in each case a path with minimum number of mistakes. First, consider the resistance from the non-cooperation equilibrium e^1 to the partial cooperation equilibrium e^2 . Any path from e^1 to e^2 must contain e_{-i}^2 (or e_{-i}^3) for every $i \in N_1$. When the cost ε_i is small enough, against a sample that contains one e_{-i}^2 and $k - 1$ non-cooperative profiles, i 's best response is to participate. Thus the path depicted in Figure 2 is a path from e^1 to e^2 .

(Figure 2 appears about here.)

⁷Strictly speaking, it is a sequence of states, rather than that of action profiles, that matters in the analysis. In practice, however, we can analyze resistances in terms of sequences of action profiles.

In this and similar figures that follow, an action by mistake is indicated by an asterisk, as 1^* . Figure 2 shows that the adaptive play can transit from e^1 to e^2 with n_1 mistakes. Therefore, $r(e^1, e^2) \leq n_1$. In fact, it is clear that $r(e^1, e^2) = n_1$ if $n_1 \geq 3$. In contrast, $r(e^1, e^2) = n_1 - 1 = 1$ if $n_1 = 2$, as in Figure 3.

(Figure 3 appears about here.)

For the resistance $r(e^1, e^3)$, it is clear that $r(e^1, e^2) < r(e^1, e^3)$, the only inequality we need to concern here regarding $r(e^1, e^3)$.

Next, consider the resistance from the partial cooperation equilibrium e^2 to the non-cooperation equilibrium e^1 . There are two types of paths to be distinguished. First, consider the path in Figure 4. On a date in the phase 1, all players $j \in N_2 \setminus \{n_1 + 1\}$ happen to make mistakes simultaneously. From the next date on, up to the point where a sufficient number of e_{-i}^3 's accumulate ($i = n_1 + 1$), the simultaneous mistakes occur consecutively. This yields a sample for i , namely the entire phase 1, that exclusively consists of e_{-i}^2 's and e_{-i}^3 's. Against phase 1, the best response of player i is to participate if

$$sa_i + (k - s)c_i \geq sd_i + (k - s)f_i,$$

where s ($k - s$, resp.) is the number of e_{-i}^3 (e_{-i}^2 , resp.) in the sample. Therefore the sufficient number of e_{-i}^3 's turns out to be at least $\alpha_i k$, where

$$\alpha_i = \left(\frac{f_i - c_i}{a_i - d_i + f_i - c_i} \right).$$

For this type of transition to happen, at least $(n_2 - 1) \lceil \alpha_i k \rceil$ mistakes are required, where $\lceil z \rceil$ is the minimum integer greater or equal to a real number z .

(Figure 4 appears about here.)

Notice that this kind of path will lead to e^1 , not necessarily to e^3 . More specifically, when $n_2 \geq 3$ the path never reaches to e^3 without further mistakes. If $n_2 = 2$, the path may reach to e^3 , as well as to e^1 . In phase 2 in Figure 4, everyone samples phase 1, and optimally responds to it. This results in $\sigma_j = 1$ for $j = 1, \dots, n_1, n_1 + 1$ and $\sigma_j = 0$ for $j = n_1 + 2, \dots, n_1 + n_2$. In phase 3, by sampling phase 2 and optimally responding to it, no one chooses to participate if

$n_2 \geq 3$ ($x = 0$). If $n_2 = 2$, only player $n_2 + 1$ participates ($x = 1$), but by now it is clear that the adaptive play is heading for the non-cooperation equilibrium.⁸

There is another type of paths that connect e^2 and e^1 . An example is given in Figure 5.

(Figure 5 appears about here.)

In phase 1, every player is given the sample that entirely consists of e^2 . Every player except $n_1 + 1$ optimally responds to it. The player $n_1 + 1$, meanwhile, continues to participate by mistake. In phase 2 and beyond, every player receives the previous phase as the sample, and optimally responds to it. In phase 2, players $i \neq n_1 + 2$ choose not to participate. The choice x of player $n_1 + 2$ depends on n_2 . If $n_2 \geq 3$, then $x = 0$. $x = 1$ when $n_2 = 2$. In any case, the path results in e^1 in phase 3. The total number of mistakes is k .

Comparing two types of paths, it follows that the number of potential free-riders, n_2 , and the sample size, k , matter. That is, the resistance $r(e^2, e^1)$ is equal to $(n_2 - 1) \lceil \alpha_i k \rceil$ if $(n_2 - 1) \lceil \alpha_i k \rceil < k$. Otherwise, it is k . Intuitively, $r(e^2, e^1) = k$ when n_2 is “large.” It is less than k when n_2 is “small.” It is worth emphasizing that the large/small distinction is relevant because we allow three or more players in N_2 .⁹ A similar consideration calls for the resistance from the full cooperation equilibrium e^3 to the non-cooperation equilibrium e^1 .

To summarize the observations, we introduce a number of definitions.

Definition 4.1. The *incentive ratio* of player $i \in N_2$ is the fraction

$$\eta_i = \frac{a_i - d_i}{f_i - d_i}.$$

The population size n_2 of N_2 is *large to exit from e^2* if

$$n_2 - 2 \geq \max_{i \in N_2} \eta_i.$$

Otherwise, n_2 is *small to exit from e^2* . Similarly, n_2 is *large to exit from e^3* if

$$n_2 - 2 \geq \max_{i \in N_2} \frac{1}{\eta_i}.$$

Otherwise, n_2 is *small to exit from e^3* .

⁸Assume $n_2 = 2$ and consider phase 3. If everyone but $n_1 + 2$ still samples phase 1, while $n_1 + 2$ samples phase 2, the adaptive play leads to e^3 .

⁹Dealing with pure coordination games in a random matching setup, as opposed to the fictitious play model considered here, Young (1998) discusses a similar point.

Ignoring the small participation cost ε_i , the incentive ratio is the ratio of “deviation losses” in e^2 and e^3 . Note that n_2 is small to exit from both e^2 and e^3 if $n_2 = 2$. An $n_2 \geq 3$ can be large to exit from both e^2 and e^3 .

Define

$$\alpha_i = \frac{f_i - c_i}{a_i - d_i + f_i - c_i} \quad \text{and} \quad \beta_i = \frac{a_i - d_i}{a_i - d_i + f_i - c_i}.$$

Set $\alpha = \min_{i \in N_2} \alpha_i$ and $\beta = \min_{i \in N_2} \beta_i$.

The next lemma collects the observations thus far. In particular, (2) and (3) confirm that the large/small distinction defined above works as desired. Proofs are given in the Appendix.

Lemma 4.1.

- (1) *If $n_1 \geq 3$, $r(e^1, e^2) = n_1$. If $n_1 = 2$, $r(e^1, e^2) = 1$. In either case, $r(e^1, e^2) < r(e^1, e^3)$.*
- (2) *If n_2 is large to exit from e^2 , then $r(e^2, e^1) = k < r(e^2, e^3)$. If n_2 is small to exit from e^2 , then*

$$(n_2 - 1)\alpha k \leq r(e^2, e^1) \leq (n_2 - 1) \lceil \alpha k \rceil.$$

In either case, $r(e^2, e^1) \leq r(e^2, e^3)$ and $r(e^2, e^1) \leq k$.

- (3) *If n_2 is large to exit from e^3 , then $r(e^3, e^1) = k$. If n_2 is small to exit from e^3 , then*

$$(n_2 - 1)\beta k \leq r(e^3, e^1) \leq (n_2 - 1) \lceil \beta k \rceil.$$

In either case, $r(e^3, e^1) \leq r(e^3, e^2)$ and $r(e^3, e^1) \leq k$.

4.3 Equilibrium Selection

We are now ready to present a number of equilibrium selection results for the group formation game with two types. Proofs are given in the Appendix.

Lemma 4.1 contains two types of evaluations of resistances. First, the resistances out of e^2 or e^3 to e^1 are evaluated in terms of k . Specifically, they are at most k , and some of them are less than k only if n_2 is small. Second, the resistances out of e^1 are independent of k , and are evaluated in terms of n_1 . Having made no assumption thus far, there is no way to compare the two types of resistances. In other words, equilibrium selection results depend on the relative magnitude of k and n_1 . If $n_1 > k$, it is more difficult for the revision process to exit from e^1 than to enter into e^1 . This leads to the following result.

Theorem 4.1. *Assume that $n_1 \geq 3$. If $n_1 > k$, then the non-cooperation equilibrium is uniquely stochastically stable.¹⁰*

As Figure 2 shows, in order for the adaptive play to escape from the non-cooperation equilibrium, simultaneous mistakes by all members of N_1 must occur at least once. The larger the size of N_1 is, the more unlikely such an event becomes. On the other hand, Figure 5 shows that a k -successive mistakes is enough for the adaptive play to enter the non-cooperation equilibrium. The smaller the sample size is, the more likely such an event becomes. As a result, if $n_1 > k$, then the adaptive play is more likely to enter into the non-cooperation equilibrium, rather than to escape from it. Hence, the non-cooperation equilibrium is stochastically stable.

The next two results together show that non-cooperation equilibrium is not stochastically stable when the sample size k is larger than the subpopulation size n_1 . In this case, the number n_2 of potential free-riders determines the selection outcome.

Theorem 4.2. *Assume that $n_1 < k$. If n_2 is large to exit from e^2 , the partial cooperation equilibrium e^2 is uniquely stochastically stable.¹¹*

In words, when the number of potential free-riders exceeds a critical level, the partial cooperation equilibrium is uniquely stable in the long-run. Recall that n_2 is large to exit from e^2 if

$$n_2 - 2 \geq \max_{i \in N_2} \frac{a_i - d_i}{f_i - d_i}.$$

The condition suggests at least two interpretations of the result. First, given the maximum incentive ratio of the potential free-riders, the theorem states that the free-riding equilibrium is the unique stable outcome when there are sufficient number of them. In other words, the full cooperation is difficult to emerge when there are many potential free-riders. Second, the theorem tells us, quite naturally, when the incentive to free-ride is sufficiently strong, the free-riding equilibrium is likely to be observed in the long-run. This is because, given the number of potential free-riders, the stronger the incentive to free-ride, the smaller the incentive ratio. Notice that if the free-riding payoff is larger than the universal cooperative payoff, that is, $f_i > a_i$, for every $i \in N_2$, the theorem applies for every $n_2 \geq 3$. Since we allow heterogeneous

¹⁰If $n_1 = 2$ and $k = 1$, then e^1 and e^2 are stable.

¹¹If $k = n_1$, one can show the following results. If n_2 is large to exit from e^2 , then the set of stochastically stable equilibria is $\{e^1, e^2\}$. If n_2 is small to exit from e^2 , then e^1 is uniquely stochastically stable.

preferences, however, some players may well have large incentive ratios. In such a case, the assumption of the theorem becomes harder to be satisfied.

It remains to consider the case in which the number of potential free-riders is smaller than the critical level. In this case, it turns out that the stochastically stable outcome is determined by a variant of risk-dominance relation (Harsanyi and Selten 1988). Before stating the result, let us introduce the risk-dominance relation relevant here.

Assume that all players in N_2 expect that the game will be played according to either the full cooperation equilibrium or the partial cooperation equilibrium, but they are not certain about which equilibrium will prevail. Suppose that each player i in N_2 expects that the partial cooperation equilibrium is played with probability t , and the full cooperation equilibrium with probability $1 - t$. If she participates in a group, she receives expected payoff $td_i + (1 - t)a_i$ (neglecting small participation costs ε_i). If she does not participate, she receives expected payoff $tf_i + (1 - t)d_i$. Then, it is optimal for her to stay at the full cooperation equilibrium if $t < \frac{\eta_i}{1 + \eta_i}$. Thus, $\min_{i \in N_2} \frac{\eta_i}{1 + \eta_i}$ can be interpreted as the maximum level of risk that all players in N_2 can take in staying at the full cooperation equilibrium. Similarly, $\min_{i \in N_2} \frac{1}{1 + \eta_i}$ can be interpreted as the maximum level of risk that all players in N_2 can take in staying at the partial cooperation equilibrium. Specifically, if

$$\min_{i \in N_2} \frac{\eta_i}{1 + \eta_i} > \min_{i \in N_2} \frac{1}{1 + \eta_i}, \quad (4.1)$$

then the full cooperation equilibrium is more “robust” than the partial cooperation equilibrium in the risk consideration. In this case, following the spirit of Harsanyi and Selten (1988), we say that the full cooperation equilibrium *risk-dominates* the partial cooperation equilibrium. The next theorem shows that it is this notion of risk-dominance that works for the criterion of stochastic stability in the group formation game.

Theorem 4.3. *Assume that k is sufficiently large and n_2 is small to exit from e^2 .*

- (1) *The full cooperation equilibrium is uniquely stochastically stable if*

$$\min_{i \in N_2} \frac{\eta_i}{1 + \eta_i} > \min_{i \in N_2} \frac{1}{1 + \eta_i}.$$

- (2) *The partial cooperation equilibrium is uniquely stochastically stable if*

$$\min_{i \in N_2} \frac{\eta_i}{1 + \eta_i} < \min_{i \in N_2} \frac{1}{1 + \eta_i}.$$

As the usual risk-dominance relation in 2×2 coordination games, ours in (4.1) can be restated in terms of deviation losses. Let η_M (η_m , resp.) be the highest (lowest, resp.) incentive ratio among all potential free-riders in N_2 . Then, the risk-dominance condition (4.1) is equivalent to

$$\frac{\eta_m}{1 + \eta_m} > \frac{1}{1 + \eta_M},$$

which can be reduced to $\eta_m \eta_M > 1$, or

$$(a_m - d_m)(a_M - d_M) > (f_m - d_m)(f_M - d_M).$$

The last inequality makes it clear that the risk-dominance here is a variant of the original version of Harsanyi and Selten (1988). In particular, the two coincide each other when $n_2 = 2$. Moreover, when the full cooperation equilibrium strictly Pareto dominates, that is, $a_i > f_i$ for every $i \in N_2$, then the former risk-dominates the latter, and thus the full cooperation equilibrium is stochastically stable.

It is now instructive to consider the following game. In the group formation game, fix the action of every player in N_1 at the participation. The resulting game is a coordination game played by potential free-riders, in which there are exactly two strict equilibria, the partial cooperation and the full cooperation. The intuition behind Theorem 4.3 is that, when n_2 is small, the stochastically stable outcome of the whole game is the same as that of the restricted coordination game. Therefore the outcome is determined by the risk-dominance relation. Contrary to the original version of Harsanyi and Selten (1988), however, the relevant risk-dominance relation involves only the maximum incentive ratio and the minimum incentive ratio. In stochastic stability analysis, only the minimum number of mistakes to upset a given equilibrium matters. As a result, the outcome is insensitive to incentive ratios of “intermediate” players.

Let us illustrate the equilibrium selection results by means of a numerical example. For simplicity, we assume that group costs are zero. That is, $v_i = u_i$. Let the sample size k be sufficiently large.

Example 4.1. Consider a seven-person prisoner’s dilemma with the player set $N = \{1, \dots, 7\}$. Figure 6 defines payoffs. For each player i , the payoff by defection is commonly given in row $u(D, h)$, where h ($= 0, 1, \dots, 6$) is the number of other cooperators. Payoffs by cooperation are heterogeneous. Player i ’s payoff by cooperation is shown in row $u_i(C, h)$. Let us assume that $0 < a_3 < \dots < a_7 < 15$.

(Figure 6 appears about here.)

In the non-cooperative equilibrium of the prisoner's dilemma, every player receives zero. Therefore, the threshold s_i of player i is $s_1 = s_2 = 2$ and $s_3 = \dots = s_7 = 7$. Thus the group formation game associated with this game has two types of players. Namely, $N_1 = \{1, 2\}$ and $N_2 = \{3, \dots, 7\}$. By Proposition 4.1, it has three strict Nash equilibria: the non-cooperation e^1 , the partial cooperation e^2 , and the full cooperation e^3 .

Since the incentive ratio η_i is given by $\eta_i = a_i/4$, $\eta_3 < \dots < \eta_7$. Let us examine how the selection outcome depends on the payoff parameters a_3 and a_7 . If the highest incentive ratio is less than or equal to $n_2 - 2 = 3$, that is, $a_7 \leq 12$, then Theorem 4.2 states that the partial cooperation equilibrium is uniquely stochastically stable. This result is independent of other payoff parameters. If the incentive ratio η_7 increases to exceed 3, then Theorem 4.3 applies. In such a case, the selection outcome is determined by the relative magnitude of the lowest incentive ratio η_3 and the highest incentive ratio η_7 . Specifically, the full cooperation equilibrium is uniquely stochastically stable if and only if $\eta_3\eta_7 > 1$, or $a_3a_7 > 16$.

The result parallels the usual risk dominance argument applied to the restricted coordination game in Figure 7. In this game, player 3 and player 7 choose either to participate or not. The game has two strict Nash equilibria, $(1, 1)$ and $(0, 0)$, which correspond to the full cooperation equilibrium and the partial cooperation equilibrium in the group formation game, respectively. The well-known result in 2×2 coordination games tells us that $(1, 1)$ risk dominates the other if and only if $a_3a_7 > 16$. Therefore the full cooperation equilibrium in the group formation game is stochastically stable if and only if $(1, 1)$ risk dominates $(0, 0)$ in the restricted coordination game.

(Figure 7 appears about here.)

In a general two-type game, the restricted coordination game takes the following form, where $m \in N_2$ is a player with the smallest incentive ratio and $M \in N_2$ is a player with the largest incentive ratio. Whenever n_2 is small to exit from e^2 , one can identify the long-run outcome of the group formation game by simply checking the risk-dominance relation in the restricted game in Figure 8.

(Figure 8 appears about here.)

5 Conclusion

We have investigated the formation and the long-run stability of voluntary groups in collective action with heterogeneous individuals. In real situations, individuals often differ in their willingness to cooperate. We have shown that the heterogeneous preferences yield a genuine multiplicity of strict Nash equilibria in the group formation game. Different levels of cooperation are possible in equilibrium, ranging from non-cooperation to universal cooperation. Typically, there exist many partial cooperation equilibria in which cooperators and free-riders co-exist. We have investigated the equilibrium selection problem by applying the stochastic evolutionary theory to examine which level of cooperation can be sustained in the long-run. The analysis of the two types model have shown that the non-cooperation equilibrium is not stable in the long-run when every individual can collect a sufficiently large number of samples from the past plays. In this case, the long-run outcome is determined by two factors, the number of potential free-riders and their incentive ratios of cooperation. If the number of potential free-riders is larger than a critical level, a partial cooperation equilibrium is uniquely stochastically stable. Otherwise, the long-run equilibrium is determined by a risk-dominance criterion, which exclusively focuses the two distinguished potential free-riders, who respectively have the highest incentive ratio and the lowest incentive ratio. The full cooperation equilibrium is uniquely stochastically stable if there exists at least one potential free-rider whose incentive to cooperate is high enough.

References

- Diermeier, D. and J.A. Van Mieghem (2000), “Spontaneous Collective Action,” CMSEMS Discussion Paper # 1302, MEDS, Northwestern University.
- Fudenberg, D. and E. Maskin (1986), “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information,” *Econometrica*, Vol. 54, 533–554.
- Goyal, S. and F. Vega-Redondo (2000) “Learning, Network Formation, and Coordination,” mimeo. University of Alicante.
- Granovetter, M. (1978), “Threshold Models of Collective Behavior,” *American Journal of Sociology*, Vol. 83, 1420–1443.

- Harsanyi, J.C. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*, Cambridge: The MIT Press.
- Kandori, M., G. Mailath, and R. Rob (1993), “Learning, Mutation and Long-Run Equilibria in Games,” *Econometrica* Vol. 61, 29–56.
- Marwell, G. and P. Oliver (1993), *The Critical Mass in Collective Action: A Micro-Social Theory*, Cambridge: Cambridge University Press.
- Okada, A. (1993), “The Possibility of Cooperation in an n -Person Prisoners’ Dilemma with Institutional Arrangements,” *Public Choice*, Vol.77, 629–656.
- Olson, M. (1965), *The Logic of Collective Action*, Cambridge: Harvard University Press.
- Ostrom, E. (1998), “A Behavioral Approach to the Rational Choice Theory of Collective Action,” *American Political Science Review* Vol. 92, 1–22.
- Palfrey R.T. and H. Rosenthal (1984), “Participation and the Provision of Discrete Public Goods: A Strategic Analysis,” *Journal of Public Economics*, Vol. 24, 171–193.
- Schelling, T.C. (1978), *Micro-Motives and Macro-Behavior*, New York: W.W. Norton.
- Taylor, M. (1987), *The Possibility of Cooperation*, New York: Cambridge University Press.
- Young, P.H. (1993), “The Evolution of Conventions,” *Econometrica* Vol. 61, 57–84.
- Young, P.H. (1998), “Conventional Contracts,” *Review of Economic Studies* Vol. 65, 776–792.

Appendix

Proofs of the results in Section 4 are given below. We start with a definition. Recall that the resistance $r(e, e')$ is a positive integer γ such that any path from an equilibrium e to another equilibrium e' contains at least γ mistakes, and that there is such a path with exactly γ mistakes. To evaluate a resistance from below, it proves useful to consider an *exiting path* from the originating equilibrium and its *first exitors*. Recall that m is the memory size of the adaptive play with mistakes.

Definition A.1. Given an equilibrium e , an *exiting path* from e is a path

$$\overbrace{e, \dots, e}^m, \sigma^1, \dots, \sigma^T$$

of action profiles¹² such that it contains a profile σ in which some player $i \in N$ plays a best

¹²See the footnotes in Section 4.

response σ_i different from e_i . For an existing path from e , a player $i^* \in N$ is called a *first exitor* if i^* plays a best response that differs from the equilibrium e for the first time during the path.

For example, a path from e^2 is an exiting path if it contains σ such that $\sigma_i = 1$ for some $i \in N_2$ and this choice is a best response. Thus in this case, the path contains a sample to which $i \in N_2$ participates optimally. Note that a first exitor need not be unique, since two or more players can optimally switch simultaneously. Take an exiting path ξ from e and let $i \in N$ be its first exitor, who chooses optimally $\sigma_i^\tau \neq e_i$ for the first time on date τ . Then for every $j \in N$ and every date $\tau' < \tau$ any action $\sigma_j^{\tau'}$ is a mistake whenever $\sigma_j^{\tau'} \neq e_j$. Let $\gamma(\xi)$ be the number of such mistakes. Set $\gamma = \min_\xi \gamma(\xi)$, where ξ ranges over the set of all exiting paths from e . Then, $r(e, e') \geq \gamma$, since any path from e to e' is an exiting path from e . Moreover, if there is a path from e to e' that contains exactly γ mistakes, then $r(e, e') = \gamma$.

Proof of Lemma 4.1.

We assume throughout that every ε_i is sufficiently small. In other words, ε_i 's are positive but practically zero. The lemmas below collectively prove Lemma 4.1.

Lemma A.1. $r(e^1, e^2) = n_1$ if $n_1 \geq 3$ and $r(e^1, e^2) = n_1 - 1$ if $n_1 = 2$.

Proof. Consider the following path starting from e^1 . From date 1 to date $k - t$, everyone plays 0. From date $k - t + 1$ to date k , every $i \in N_1$ by mistake plays 1. Meanwhile, every $j \in N_2$ plays 0. On date $k + 1$, everyone plays a best response against the most recent k dates. The action 1 is a best response for $i \in N_1$ if and only if

$$ta_i + (k - t)c_i \geq kd_i \quad \text{or} \quad t \geq \frac{k(d_i - c_i)}{a_i - c_i} = \frac{k\varepsilon_i}{a_i - d_i + \varepsilon_i}.$$

For small enough ε_i , $t = 1$ satisfies the constraint. Therefore every $i \in N_1$ can optimally play 1 on date $k + 1$ on. Meanwhile, every $j \in N_2$ optimally continues to choose 0. Thus the path in Figure 2 is indeed a path from e^1 to e^2 . This proves that $r(e^1, e^2) \leq n_1$.

By a similar argument, the path in Figure 3 is a path from e^1 to e^2 when $n_1 = 2$. Thus, $r(e^1, e^2) \leq n_1 - 1$. Since a resistance cannot be zero, $r(e^1, e^2) = n_1 - 1$. It remains to show that $r(e^1, e^2) = n_1$ if $n_1 \geq 3$. This follows from the following Claim.

Claim Assume that $n_1 \geq 3$. Consider a path from e^1 to σ , where σ is an action profile in which $\sigma_i = 1$ for every $i \in N_1$. If none of the 1's chosen by $i \in N_1$ in σ is a mistake, then this path contains at least n_1 mistakes by members of N_1 .

For each $i \in N_1$, let τ_i be the date on which i chooses 1 as a best response for the first time during the path. Every 1 chosen by i prior to date τ_i is a mistake. Prior to date τ_i , there must be a date τ_i^* on which either e_{-i}^2 or e_{-i}^3 is played. If there are more than one such dates, let τ_i^* be the earliest date. Assume that they appear as $\tau_1^* \leq \dots \leq \tau_{n_1}^*$. If $\tau_1^* = \tau_2^*$, then the play of this date is either e^2 or e^3 , every 1 in which is a mistake. Thus the path contains at least n_1 mistakes. If $\tau_1^* < \tau_2^*$, then there are at least $n_1 - 1$ mistakes on date τ_1^* . On date τ_2^* , a 1 by player 1 may be a best response, but a 1 by player 3, who exists since $n_1 \geq 3$, must be a mistake. Thus the path contains at least n_1 mistakes. This concludes the proof of the claim. \square

Lemma A.2. $r(e^1, e^2) < r(e^1, e^3)$.

Proof. It follows from Lemma A.1, the Claim in its proof, and the fact that any path from e^1 to e^3 contains at least one mistake by a member of N_1 and at least one mistake by a member of N_2 . \square

Lemma 4.1.(1) follows from Lemmas A.1 and A.2.

Lemma A.3. *There is a path from e^2 to e^1 with exactly k mistakes. That is, $r(e^2, e^1) \leq k$.*

Proof. Such a path is given in Figure 5. See the text following Figure 5. \square

Lemma A.4. *Any exiting path from e^2 with a first exitor in N_1 contains at least k mistakes.*

Proof. The proof of Lemma A.1 shows that for $i \in N_1$ to optimally choose 0, the sample must contain no e_{-i}^2 . That is, a first exitor $i \in N_1$ must have a sample in which every action profile contains at least one mistake. \square

In the next lemma, we are going to setup an integer program. In this program, t and s are nonnegative integer variables that satisfy $t + s \leq k$, where k is the sample size of the adaptive play. The program relates to the mistake counting argument as follows. Take an exiting path from e^2 with a first exitor $i \in N$. The solution gives us the number of mistakes the path must contain prior to the date on which i optimally chooses $\sigma_i \neq e_i^2$ for the first time during the path.

Lemma A.5. *If n_2 is large to exit from e^2 , then any exiting path from e^2 with a first exitor in N_2 contains more than k mistakes.*

Proof. Take any exiting path from e^2 with a first exitor in N_2 . Recall that a player $i \in N_2$ optimally plays action 1 only against e_{-i}^3 . For any other action profile, 0 is the unique best response. Let the sample, to which the first exitor $i \in N_2$ optimally responds by action 1, contain the action profile e_{-i}^3 for t times, e_{-i}^2 for $k - t - s$ times, and s others. Since i is a first exitor, each 1 in e_{-i}^3 played by a member in N_2 is a mistake (there are $n_2 - 1$ of them), and each of the “other s ” profiles contains at least one mistake. One can evaluate the minimum number of mistakes by solving the following linear program. Note that 1 is a best response against this sample if and only if the constraint of the following program is satisfied:

$$\begin{aligned} \min (n_2 - 1)t + s, \quad \text{subject to} \quad & ta_i + (k - t)c_i \geq (k - t - s)f_i + (t + s)d_i, \\ & t \geq 0, \quad s \geq 0, \quad k \geq t + s. \end{aligned} \tag{P.1}$$

The exact value of the minimum number of mistakes is given by program (P.1) with integer constraints. Ignoring integer constraint, the optimal value of (P.1) is less than or equal to the minimum number of mistakes. Thus it suffices to show that the optimal value of (P.1) exceeds k . The main constraint in (P.1) is equivalent to

$$s \geq \frac{k(f_i - c_i)}{f_i - d_i} - \left(\frac{a_i - d_i + f_i - c_i}{f_i - d_i} \right) t.$$

Draw a horizontal t axis and a vertical s axis. See Figure 9. In this coordinate, the boundary of constraint (P.1) is a line that has a negative slope steeper than -1 and its intercept on s axis is above k . On the other hand, the slope of the objective function is $-(n_2 - 1)$. Note that when the objective function passes through an optimal solution of (P.1), its intercept on s axis gives the optimum value of (P.1).

(Figure 9 appears about here.)

By assumption, $n_2 - 2 \geq \eta_i$. Let us assume that $n_2 - 2 > \eta_i$. We skip the proof of the nongeneric case $n_2 - 2 = \eta_i$. Since ε_i is small, we have

$$n_2 - 1 \geq \frac{a_i - d_i + f_i - c_i}{f_i - d_i},$$

which means that the objective function is (weakly) steeper than the constraint boundary. Thus $(\underline{t}, \underline{s})$ in Figure 9 is an optimum solution. Clearly, when the objective function passes through $(\underline{t}, \underline{s})$, the intercept is above \bar{s} , which in turn strictly exceeds k . Therefore the optimal value is greater than k . \square

Lemmas A.3, A.4, and A.5 show that $r(e^2, e^1) = k \leq r(e^2, e^3)$ if n_2 is large to exit from e^2 .

Lemma A.6. *If n_2 is large to exit from e^2 , then $r(e^2, e^3) > k$.*

Proof. Assume that n_2 is large to exit from e^2 and take any path from e^2 to e^3 . Let $i^* \in N_2$ be a first player in N_2 who chooses 1 as a best response during the path. Let τ be the date on which i^* optimally chooses 1 for the first time. If $i^* \in N_2$ is a first exitor of this path, then it contains more than k mistakes by Lemma A.5. Thus we can assume that $i^{**} \in N_1$ is a first exitor, but i^* is not. Similarly to the Claim in Lemma A.1, one can verify in this case that at least $k + n_2 - 2$ mistakes has been made prior to date τ . Note that $n_2 \geq 3$, Since n_2 is large to exit from e^2 . \square

Lemma A.7. *Assume that n_2 is small to exit from e^2 . Then any exiting path from e^2 with a first exitor in N_2 contains at least $(n_2 - 1)t^* + s^*$ mistakes, where t^* and s^* are nonnegative integers such that*

$$(n_2 - 1)\alpha k \leq (n_2 - 1)t^* + s^* \leq (n_2 - 1) \lceil \alpha k \rceil.$$

In addition, there is a path from e^2 to e^1 with exactly $(n_2 - 1)t^ + s^*$ mistakes.*

Proof. Assume that n_2 is small to exit from e^2 . Then for some $i \in N_2$,

$$n_2 - 1 < \frac{a_i - d_i + f_i - c_i}{f_i - d_i}. \quad (\dagger)$$

Take an exiting path from e^2 with first exitor in $i \in N_2$. The number of mistakes such a path contains prior to the date on which i begins to choose 1 optimally is given by the optimal value of program (P.1) (in the proof of Lemma A.5). By (\dagger) , the slope of the objective function in (P.1) is flatter than that of the constraint boundary. Ignoring integer constraints, the optimum solution is $(\alpha_i k, 0)$ in Figure 9. Thus the optimal value with integer constraint is at least $(n_2 - 1)\alpha_i k$. Rounding $\alpha_i k$ gives $\lceil \alpha_i k \rceil$. Since $\lceil \alpha_i k \rceil$ is an integer, the optimal value with integer constraint is at most $(n_2 - 1) \lceil \alpha_i k \rceil$.

Note that $(t, s) = (\lceil \alpha_i k \rceil, 0)$ need not be the optimal solution of (P.1) with integer constraints. For each $i \in N_2$ such that (\dagger) holds, let (t^i, s^i) be an optimal solution of (P.1) with integer constraints. Let $(t^*, s^*) = (t^{i^*}, s^{i^*})$ be a minimizer of $(n_2 - 1)t^i + s^i$ over $i \in N_2$ such that (\dagger) holds. We know the following:

$$(1) \quad (n_2 - 1)\alpha k \leq (n_2 - 1)t^* + s^* \leq (n_2 - 1) \lceil \alpha k \rceil.$$

(2) Against the sample that contains $e_{-i^*}^3$ for t^* times, $e_{-i^*}^2$ for $k - t^* - s^*$ times, and s^* others, the best response for i^* is 1.

(3) Any exiting path from e^2 with a first exitor in N_2 contains at least $(n_2 - 1)t^* + s^*$ mistakes.

To complete the proof, it suffices to construct a path from e^2 to e^1 with exactly $(n_2 - 1)t^* + s^*$ mistakes. Such a path is shown Figure 10, in which $i^* = n_1 + 1$. In every phase, each player receives the previous phase as her sample. As in Figure 5, the value of x depends on whether $n_2 \geq 3$ or not. \square

In Lemma A.7, it may or may not be the case that $(n_2 - 1)t^* + s^* < k$. In either case, however, Lemmas A.3, A.4, and A.7 prove Lemma 4.1.(2) when n_2 is small to exit from e^2 .

Finally, a series of arguments analogous to Lemma A.3 to A.7 shows Lemma 4.1.(3). When n_2 is small to exit from e^3 , the relevant program is

$$\begin{aligned} \min (n_2 - 1)t + s, \quad \text{subject to} \quad & tf_i + (k - t)d_i \geq (k - t - s)a_i + (t + s)c_i, \\ & t \geq 0, \quad s \geq 0, \quad k \geq t + s. \end{aligned}$$

Proofs of Equilibrium Selection Results.

The *stochastic potential* $\rho(T)$ of a tree T is defined to be the sum of the resistances of edges that are contained in T . For example, the stochastic potential of tree T_1 in Figure 1 is $\rho(T_1) = r(e^2, e^1) + r(e^3, e^2)$. Let us say that a tree T (weakly, resp.) *dominates* another tree T' if $\rho(T) < \rho(T')$ ($\rho(T) \leq \rho(T')$, resp.).

Lemma B.1. *Neither T_6 nor T_9 is a minimum tree. Moreover, T_1, T_3, T_4 are weakly dominated trees.*

Proof. By Lemma 4.1.(1), $r(e^1, e^2) < r(e^1, e^3)$. Thus T_4 and T_7 dominate T_6 and T_9 , respectively. By Lemma 4.1.(2), $r(e^2, e^1) \leq r(e^2, e^3)$. This implies that T_2 weakly dominates T_3 . By Lemma 4.1.(3), $r(e^3, e^1) \leq r(e^3, e^2)$. Therefore T_2 and T_5 weakly dominate T_1 and T_4 , respectively. \square

Proof of Theorem 4.1. By Lemma 4.1 and the assumption that $n_1 > k$,

$$\max\{r(e^2, e^1), r(e^3, e^1)\} < r(e^1, e^2) \leq r(e^1, e^3).$$

By Lemma B.1, it suffices to show that T_2 dominates T_7 , T_8 , and T_5 . First, $r(e^2, e^1) \leq r(e^2, e^3)$ and $r(e^3, e^1) < r(e^1, e^2)$ imply that T_2 dominates T_7 . Second, $r(e^3, e^1) < r(e^1, e^3)$ implies that T_2 dominates T_8 . Third, $r(e^2, e^1) < r(e^1, e^2)$ implies that T_2 dominates T_5 . \square

Proof of Theorem 4.2. By Lemma B.1, it suffices to show that T_5 dominates T_7 , T_8 , and T_2 . It follows from Lemma 4.1.(3) that $r(e^3, e^1) \leq k$. If n_2 is large to exit from e^2 , then $r(e^2, e^3) > k$ by Lemma 4.1.(2). Thus $r(e^2, e^3) > r(e^3, e^1)$. Therefore T_5 dominates T_7 . Likewise, we have $r(e^2, e^1) = k$ by Lemma 4.1.(2). Thus $r(e^2, e^1) \geq r(e^3, e^1)$. This inequality and Lemma 4.1.(1) together imply that

$$r(e^2, e^1) + r(e^1, e^3) > r(e^3, e^1) + r(e^1, e^2).$$

Thus T_5 dominates T_8 . Finally, $r(e^2, e^1) = k > n_1 \geq r(e^1, e^2)$ implies that T_5 dominates T_2 . \square

To prove Theorem 4.3, we make use of the following lemma. Proofs are skipped, but they are available from the authors on request.

Lemma B.2.

- (1) If n_2 is small to exit from e^2 , then $k > (n_2 - 1) \lceil \alpha k \rceil + n_2$ for sufficiently large k .
- (2) $r(e^1, e^3) \leq r(e^1, e^2) + n_2$.
- (3) $r(e^3, e^1) \leq (n_2 - 1) \lceil \beta k \rceil$.

Proof of Theorem 5.3. First of all, by Lemma 4.1.(1) and (2), $r(e^1, e^2) \leq n_1$ and $(n_2 - 1)\alpha k \leq r(e^2, e^1)$, respectively. It is clear that $r(e^1, e^2) < r(e^2, e^1)$ for sufficiently large k so that T_5 dominates T_2 .

For (1), assume that $\min_i \eta_i / (1 + \eta_i) > \min_i 1 / (1 + \eta_i)$, or equivalently

$$\min_{i \in N_2} \frac{a_i - d_i}{a_i - d_i + f_i - d_i} > \min_{i \in N_2} \frac{f_i - d_i}{a_i - d_i + f_i - d_i}.$$

For sufficiently large k and sufficiently small ε_i , we have

$$\min_{i \in N_2} \frac{a_i - d_i}{a_i - d_i + f_i - c_i} > \min_{i \in N_2} \frac{f_i - c_i}{a_i - d_i + f_i - c_i} + \frac{1}{k} \left(1 + \frac{n_2}{n_2 - 1} \right),$$

where $c_i = d_i - \varepsilon_i$. This implies $\beta > \alpha + (1 + \frac{n_2}{n_2 - 1})/k$, and thus $\beta k > \lceil \alpha k \rceil + n_2 / (n_2 - 1)$. We are going to show that this inequality and Lemma B.2.(1) constitute a sufficient condition for e^3 to be stochastically stable.

Since n_2 is small to exit from e^2 , $r(e^2, e^1) \leq (n_2 - 1) \lceil \alpha k \rceil$ by Lemma 4.1.(2). On the other hand, $r(e^1, e^3) \leq r(e^1, e^2) + n_2$ by Lemma B.2.(2). Therefore

$$\rho(T_8) \leq r(e^1, e^2) + (n_2 - 1) \lceil \alpha k \rceil + n_2.$$

By Lemma B.1, it suffices to show that T_8 dominates T_5 . Assume first that n_2 is small to exit from e^3 . Then by Lemma 4.1.(3), $r(e^3, e^1) \geq (n_2 - 1)\beta k$. Thus

$$\rho(T_5) \geq r(e^1, e^2) + (n_2 - 1)\beta k.$$

Therefore e^3 is uniquely stochastically stable if

$$\begin{aligned} \rho(T_5) - \rho(T_8) &\geq r(e^1, e^2) + (n_2 - 1)\beta k - (r(e^1, e^2) + (n_2 - 1) \lceil \alpha k \rceil + n_2) \\ &= (n_2 - 1)\beta k - (n_2 - 1) \lceil \alpha k \rceil - n_2 > 0. \end{aligned}$$

The last inequality is equivalent to $\beta k > \lceil \alpha k \rceil + n_2/(n_2 - 1)$. Assume next that n_2 is large to exit from e^3 . Then by Lemma 4.1.(3), $r(e^3, e^1) = k$. Similarly to the preceding case, e^3 is uniquely stochastically stable if $k - (n_2 - 1) \lceil \alpha k \rceil - n_2 > 0$, which is equivalent to Lemma B.2.(1).

For (2), assume that $\min_i 1/(1 + \eta_i) > \min_i \eta_i/(1 + \eta_i)$. Then, similarly to (1), $\alpha k > \lceil \beta k \rceil$ for small ε_i and large k . Thus it suffices to show that the last inequality is a sufficient condition for e^2 to be stochastically stable.

By Lemma 4.1.(3), $r(e^3, e^1) \leq (n_2 - 1) \lceil \beta k \rceil$. Thus

$$\rho(T_5) \leq r(e^1, e^2) + (n_2 - 1) \lceil \beta k \rceil.$$

By Lemma B.1, it suffices to show that T_5 dominates T_7 and T_8 . Since n_2 is small to exit from e^2 , Lemma 4.1.(2) implies that $r(e^2, e^j) \geq (n_2 - 1)\alpha k$ for $j = 1, 3$. On the other hand, $r(e^1, e^3) > r(e^1, e^2)$ by Lemma 4.1.(1). Thus

$$\min\{\rho(T_7), \rho(T_8)\} \geq r(e^1, e^2) + (n_2 - 1)\alpha k.$$

Therefore e^2 is uniquely stochastically stable if

$$\begin{aligned} \min\{\rho(T_7), \rho(T_8)\} - \rho(T_5) &\geq r(e^1, e^2) + (n_2 - 1)\alpha k - r(e^1, e^2) - (n_2 - 1) \lceil \beta k \rceil \\ &= (n_2 - 1)(\alpha k - \lceil \beta k \rceil) > 0. \end{aligned} \quad \square$$

Figures

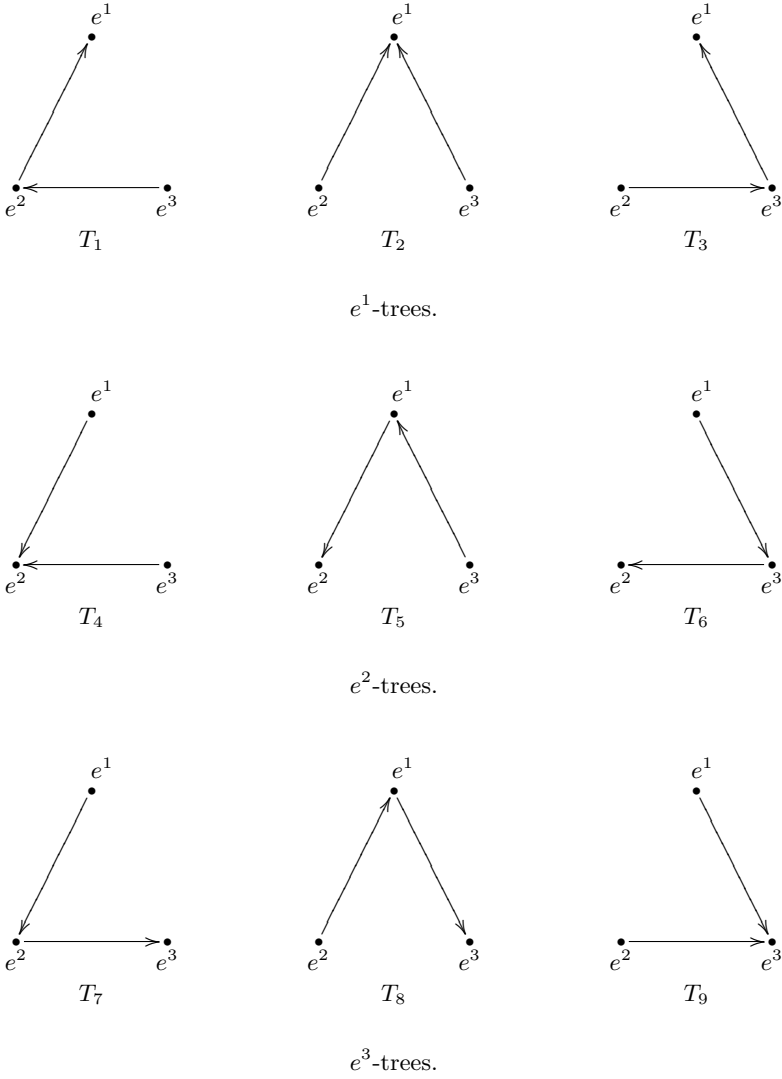


Figure 1: Trees in the group formation game.

	m				k			
σ_1	0	...	0	1*	1	...	1	
\vdots	\vdots	...	\vdots	\vdots	\vdots	...	\vdots	
σ_{n_1}	0	...	0	1*	1	...	1	
σ_{n_1+1}	0	...	0	0	0	...	0	
\vdots	\vdots	...	\vdots	\vdots	\vdots	...	\vdots	
$\sigma_{n_1+n_2}$	0	...	0	0	0	...	0	

Figure 2: A path from e^1 to e^2 with n_1 mistakes.

	m				k			
σ_1	0	...	0	1*	0	1	...	1
σ_2	0	...	0	0	1	1	...	1
σ_{2+1}	0	...	0	0	0	0	...	0
\vdots	\vdots	...	\vdots	\vdots	\vdots	\vdots	...	\vdots
σ_{2+n_2}	0	...	0	0	0	0	...	0

Figure 3: A path from e^1 to e^2 when $n_1 = 2$.

	Phase 0			Phase 1				Phase 2			Phase 3				
	m			$k - \lceil \alpha_i k \rceil$		$\lceil \alpha_i k \rceil$		k			k				
σ_1	1	...	1	1	...	1	1	...	1	1	...	1	0	...	0
\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots
σ_{n_1}	1	...	1	1	...	1	1	...	1	1	...	1	0	...	0
σ_{n_1+1}	0	...	0	0	...	0	0	...	0	1	...	1	0	...	0
σ_{n_1+2}	0	...	0	0	...	0	1*	...	1*	0	...	0	x	...	x
σ_{n_1+3}	0	...	0	0	...	0	1*	...	1*	0	...	0	0	...	0
\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots
$\sigma_{n_1+n_2}$	0	...	0	0	...	0	1*	...	1*	0	...	0	0	...	0

Figure 4: An exiting path from e^2 .

	Phase 0 m			Phase 1 k			Phase 2 k			Phase 3 k		
σ_1	1	...	1	1	...	1	0	...	0	0	...	0
\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots
σ_{n_1}	1	...	1	1	...	1	0	...	0	0	...	0
σ_{n_1+1}	0	...	0	1*	...	1*	0	...	0	0	...	0
σ_{n_1+2}	0	...	0	0	...	0	x	...	x	0	...	0
σ_{n_1+3}	0	...	0	0	...	0	0	...	0	0	...	0
\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots
$\sigma_{n_1+n_2}$	0	...	0	0	...	0	0	...	0	0	...	0

Figure 5: A path from e^2 to e^1 with exactly k mistakes.

h	0	1	2	3	4	5	6
$u(D, h)$	0	2	4	6	8	10	15
$u_1(C, h)$	-1	1	2	3	4	5	6
$u_2(C, h)$	-1	1	2	3	4	5	6
$u_3(C, h)$	-6	-5	-4	-3	-2	-1	a_3
$u_4(C, h)$	-7	-6	-5	-4	-3	-2	a_4
$u_5(C, h)$	-8	-7	-6	-5	-4	-3	a_5
$u_6(C, h)$	-9	-8	-7	-6	-5	-4	a_6
$u_7(C, h)$	-10	-9	-8	-7	-6	-5	a_7

Figure 6: A seven-person prisoner's dilemma.

	1	0
1	a_3, a_7	0, 0
0	0, 0	4, 4

Figure 7: The restricted coordination game.

	1	0
1	$u_m(C, n-1), u_M(C, n-1)$	$u_m(D, 0), u_M(D, 0)$
0	$u_m(D, 0), u_M(D, 0)$	$u_m(D, n_1), u_M(D, n_1)$

Figure 8: The restricted coordination game.

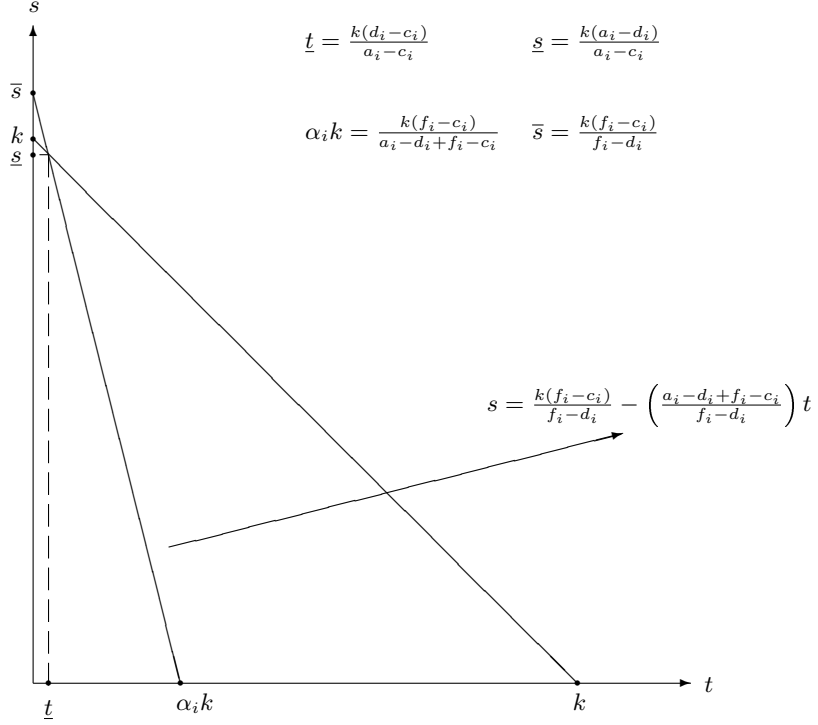


Figure 9: Program (P.1).

	Phase 0			Phase 1				Phase 2			Phase 3							
	m			$k - t^* - s^*$		s^*		t^*			k		k					
σ_1	1	...	1	1	...	1	1	...	1	1	...	1	1	...	1	0	...	0
\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots
σ_{n_1}	1	...	1	1	...	1	1	...	1	1	...	1	1	...	1	0	...	0
σ_{n_1+1}	0	...	0	0	...	0	0	...	0	0	...	0	1	...	1	0	...	0
σ_{n_1+2}	0	...	0	0	...	0	1*	...	1*	1*	...	1*	0	...	0	x	...	x
σ_{n_1+3}	0	...	0	0	...	0	0	...	0	1*	...	1*	0	...	0	0	...	0
\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots	\vdots	...	\vdots
$\sigma_{n_1+n_2}$	0	...	0	0	...	0	0	...	0	1*	...	1*	0	...	0	0	...	0

Figure 10: A path from e^2 to e^1 when n_2 is small to exit from e^2 .