Three-person game facilitates indirect reciprocity under image scoring

Shinsuke Suzuki*, Eizo Akiyama

Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-8573, Japan

Abstract

Reputation building plays an important role in the evolution of reciprocal altruism when the same individuals do not interact repeatedly because, by referring to reputation, a reciprocator can know which partners are cooperative and can reciprocate with a cooperator. This reciprocity based on reputation is called *indirect reciprocity*. Previous studies of indirect reciprocity have focused only on *two*-person games in which only two individuals participate in a single interaction, and have claimed that indirectly reciprocal cooperation cannot be established under *image scoring* reputation criterion where the reputation of an individual who has cooperated (defected) becomes good (bad). In this study, we specifically examine *three*-person games, and reveal that indirectly reciprocal cooperation can be formed and maintained stably, even under image scoring, by a *nucleus shield* mechanism. In the nucleus shield, reciprocators are a shield that keeps out unconditional defectors, whereas unconditional cooperators are the backbone of cooperation that retains a good reputation among the population.
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Evolution of cooperation; Indirect reciprocity; Reputation; Prisoner's dilemma game; *n*-person game

Indirect reciprocity based on social reputation has been considered as important in the evolution of cooperation when same individuals do not interact repeatedly (Alexander, 1987). Under indirect reciprocity, an individual who has cooperated obtains returns indirectly from someone else, who knows through social reputation that she is cooperative, in the community (Nowak and Sigmund, 2005).

Nowak and Sigmund (1998a,b) have formalized a mathematical model of indirect reciprocity as an evolutionary two-person giving game where the reputation of an opponent affects the decision-making process. In their model, pairs of individuals interact only a few times and all individuals are informed about their partners' reputations. Moreover, in this model, *image scoring* is adopted as a means to attach reputation. Under image scoring, those who cooperated (defected) in the previous interaction become associated with a good (bad) reputation. Especially, Nowak and Sigmund (1998b) have shown that, under image scoring, an indirectly reciprocal strategy, called *discriminating strategy* (DIS), which posits coopera-

tion only with opponents who have good reputations, is not an evolutionarily stable strategy (ESS) but persistent in a population consisting of DIS, unconditionally defective strategy (ALLD) and unconditionally cooperative strategy (ALLC). (Note that this model does not include error in implementation, i.e., an individual who intends to cooperate never fails to cooperate.)

However, it has been shown that, in the two-person giving game that includes error in implementation, DIS is neither persistent nor an ESS under image scoring (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2004, 2007; Brandt and Sigmund, 2004, 2006).[1] The reason DIS is not an ESS is as follows: (1) DIS strategists hurt each other in response to erroneous defections; (2) on the contrary, ALLC strategists always intend to cooperate; (3) therefore ALLC strategists do not hurt others and maintain their good reputation, never to be hurt by DIS strategists except erroneous defections; and (4) consequently, in the population mostly consisting of DIS, the fitness of DIS is less than ALLC, so the DIS population

---

*Corresponding author. Tel.: +81 298 53 5571.
E-mail address:* suzuki92@sk.tsukuba.ac.jp (S. Suzuki).

[1]In the absence of error, DIS is clearly not an ESS because ALLC is alternatively the best reply to DIS.

can be invaded by *ALLC*. Moreover, the fact that, in the presence of error in implementation, *DIS* is not persistent under image scoring has been shown by Panchanathan and Boyd (2003). They have demonstrated that, considering neutral drift and perturbations of replication to decrease *DIS*, the population of *DIS* invaded by *ALLC* is eventually taken over by *ALLD*. In summary, in the *two*-person giving game including error in implementation, *DIS* cannot evolve under image scoring. Many studies (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2004, 2007; Brandt and Sigmund, 2004, 2006; Takahashi and Mashima, 2006; Leimar and Hammerstein, 2001) have claimed that the evolution of indirect reciprocity requires a more complicated reputation criterion than image scoring.

Note that the above studies of indirect reciprocity have presumed dyadic interactions (i.e., two-person games). However, in the real world, more than two individuals often take part in a single interaction, e.g., sustainable use of common-pool resources and predator inspection in a school of fish, etc. (Hardin, 1968; Ostrom et al., 1999; Dugatkin, 1990). Therefore, we believe that not only two-person games but also $n(>2)$-person games (Boyd and Richerson, 1988; Eriksson and Lindgren, 2005; Joshi, 1987; Hauert et al., 2002, 2006) should be considered as models of interactions in human societies or ecosystems.

The evolution of indirect reciprocity in *n*-person games has been investigated by Suzuki and Akiyama (2007, submitted for publication), who have shown that, in *n*-person games, *DIS* can be an ESS under image scoring. However, regarding the evolutionary dynamics of indirect reciprocity in *n*-person games, they have investigated a population comprising only *DIS* and *ALLD*. Few studies have analyzed the evolutionary dynamics in a population that also includes *ALLC*, though it has been shown that *ALLC* plays an important role in two-person games (Panchanathan and Boyd, 2003). In the present study, expanding the model in Panchanathan and Boyd (2003), we analyze the evolutionary dynamics of indirect reciprocity under image scoring in *three*-person games in a population with *ALLC*, *ALLD* and *DIS*.[2] Moreover, we, as Panchanathan and Boyd (2003), Ohtsuki and Iwasa (2004, 2007), Brandt and Sigmund (2006), Takahashi and Mashima (2006), address the case in the presence of implementation error because, in our daily life, we sometimes make mistakes in implementation. The analyses reveal that indirect reciprocity can be formed and maintained stably under image scoring in three person games, although it has been shown that indirect reciprocity cannot in two-person games. Furthermore, in three-person games, the indirectly reciprocal cooperation is maintained by a *nucleus shield* mechanism (Lomborg, 1996). In the nucleus shield, *DIS* coexists with *ALLC* and *DIS* is the shield that keeps out *ALLD*, whereas *ALLC* is the backbone of

cooperation that maintains a good reputation among the population.

1. Evolutionary phenomena of indirect reciprocity in three-person games

Consider a population comprising infinitely numerous individuals. Each individual in the population has a *reputation* that is either *G* (*good*) or *B* (*bad*).

Each *generation* includes a number of *rounds*. After the first round, each subsequent round occurs with probability w ($0 \ll w < 1$), i.e., the expected value of the number of rounds in a generation is $1/(1-w)$.

In each round, all individuals are classified randomly into groups, each comprising three individuals; subsequently, they play a three-person prisoner's dilemma game in each group. In this game, each individual chooses either to "cooperate (C)" or "defect (D)". In this study, we assume that the payoffs for a cooperator, $V(C|k)$, and that for a defector, $V(D|k)$, where k is the number of opponents cooperating in the group, are calculated as the average of the payoffs against the *two* opponents in the *two*-person prisoner's dilemma game whose payoff is given in Table 1:

$$V(C|k) = \frac{k}{2}b - c, \tag{1}$$

$$V(D|k) = \frac{k}{2}b, \tag{2}$$

where $b > c > 0$. This form of payoff function is a natural expansion of the two-person prisoner's dilemma or giving game, which has been used also in several studies (Joshi, 1987; Eriksson and Lindgren, 2005; Lindgren and Johansson, 2001).

Moreover, implementation error is introduced with the parameter ε ($0 < \varepsilon \ll 1$). With probability ε, an individual who intends to cooperate fails to cooperate because of a mistake, etc.[3] In other words, an individual who intends to cooperate succeeds in cooperation with probability $\hat{\varepsilon} = 1 - \varepsilon$ ($0 \ll \hat{\varepsilon} < 1$). In this study, we mainly use the probability of success: $\hat{\varepsilon}$.

In this model, the reputation of opponents affects the decision-making process. What is the mechanism for

Table 1
Payoff of the *two*-person prisoner's dilemma game ($b > c > 0$)

		Player 2	
		C	D
Player 1	C	$(b-c, b-c)$	$(-c, b)$
	D	$(b, -c)$	$(0, 0)$

---

[2]We confirmed that the results in three-person games do not essentially differ qualitatively from those in more than three-person games.

[3]As in Panchanathan and Boyd (2003), Fishman (2003), we do not consider errors that foster unintentional cooperation, i.e., an individual who intends to defect never fails to defect. Furthermore, objective or subjective perception errors (Ohtsuki and Iwasa, 2004; Takahashi and Mashima, 2006) are not considered.

formation of reputation among individuals? For this study, we adopt "*image scoring*" as a *reputation criterion*, which prescribes how to judge the reputation of others based on the others' past actions. Under image scoring that was first used in Nowak and Sigmund (1998a, b), the reputation of each individual is G at the beginning of each generation. Moreover, the reputation of an individual who has defected becomes B and that of an individual who has cooperated becomes G. Image scoring is a simple reputation criterion that requires knowledge only of a past action of an opponent (first-order information) (cf. the standing reputation criterion given in Leimar and Hammerstein (2001) and Panchanathan and Boyd (2003) requires the second-order information). Furthermore, it has been reported that image scoring is widely used in the real world (Wedekind and Milinski, 2000; Milinski et al., 2001; Bshary, 2002; Bshary and Grutter, 2006).

As in some previous studies (Nowak and Sigmund, 1998b; Panchanathan and Boyd, 2003; Brandt and Sigmund, 2006; Ohtsuki and Iwasa, 2007), we consider three strategies: $ALLC$ who always cooperates, an $ALLD$ who never cooperates, and DIS who cooperates only when the other two partners in the group have reputation G. The frequencies of these strategies are denoted respectively as x_1, x_2, and x_3.

To investigate the evolution of the three strategies under the influence of natural selection, we use replicator dynamics (Taylor and Jonker, 1978)

$$\dot{x}_i = x_i(f_i - \bar{f}) \qquad (3)$$

on the invariant simplex $S_3 = \{\mathbf{x} = (x_1, x_2, x_3) \in R^3 : x_i \geqslant 0, \sum x_i = 1\}$. Here, f_i represents the fitness for strategy i ($i = 1, 2, 3$) and $\bar{f}$ is the average fitness in the population (the derivation of the fitness for each strategy is described in Appendix A).

2. Evolutionary stability

First, we discuss the evolutionary stability of DIS. We show, in Fig. 1, the parameter space for the payoff function and for the probability of the subsequent round, where DIS is an ESS in this three-person game. Region II in Fig. 1 is the parameter space, given by numerical calculation, in which $f_3 > f_1$ and $f_3 > f_2$ when $x_3 = 1$ and $x_1 = x_2 = 0$, i.e., DIS is an ESS. The figure shows that, between a small and a large cost-to-benefit ratio of cooperation, there exists a range of the medium ratio where DIS is an ESS. Specifically, if the cost-to-benefit ratio is large, DIS cannot resist invasion by $ALLD$; moreover, if the ratio is extremely small, DIS cannot resist invasion by $ALLC$. Under the medium cost-to-benefit ratio of cooperation, DIS is an ESS. Moreover, the range within which DIS is an ESS becomes larger as the probability that each of the subsequent rounds occur, w, increases.

In summary, in three-person games, DIS can be an ESS under image scoring. Note that it has already been shown
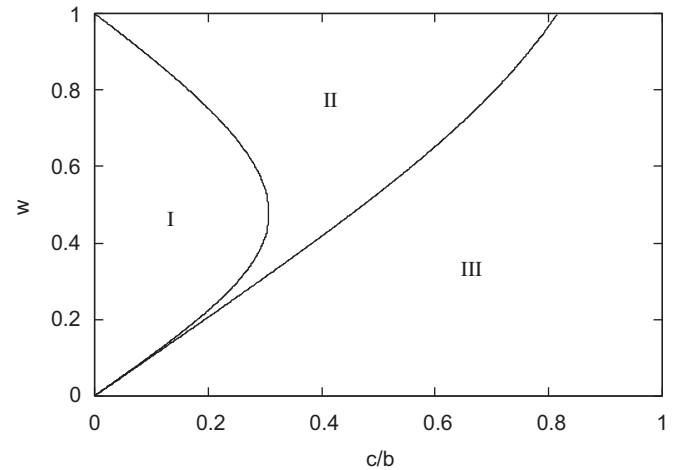

Fig. 1. Regions in $(c/b, w)$ space where DIS is an ESS ($\hat{\varepsilon} = 0.99$). Region I: DIS cannot resist invasion by $ALLC$. Region II: DIS is an ESS. Region III: DIS cannot resist invasion by $ALLD$.

that DIS is never an ESS in two-person games (Panchanathan and Boyd, 2003).

Furthermore, clearly, $ALLD$ is an ESS but $ALLC$ is not.

3. Evolutionary dynamics

Here, we show the evolutionary dynamics of the frequency of the three strategies. Moreover, from this point in our discussion, we fix w and $\hat{\varepsilon}$, respectively, as 0.95 and 0.99. (We confirmed that the overall results do not change essentially as far as $0 \ll w < 1$ and $0 \ll \hat{\varepsilon} < 1$.)

3.1. Evolutionary dynamics along the ALLD-DIS edge

First, we show the dynamics along the $ALLD$-DIS edge ($\{\mathbf{x} : x_1 = 0\}$) in the simplex S_3. On this edge, we find two stable equilibria by numerical calculation: one at the $ALLD$ corner and the other at the DIS corner; we also find one unstable polymorphic equilibrium at a point between the two corners, which we denote as F_{23}. The evolutionary dynamics on the edge are illustrated in Fig. 2(a). As that figure shows, if the cost to benefit ratio of cooperation is large (about $c/b > 0.8$), there is no internal equilibrium and the evolutionary dynamics always converge to the $ALLD$ corner. On the other hand, if c/b is sufficiently small (about $c/b < 0.8$), an unstable equilibrium exists at F_{23}. In this case, the evolutionary dynamics converge to the DIS corner when the initial value of x_3 is greater than F_{23}, and converges to the $ALLD$ corner otherwise. In other words, F_{23} is a threshold frequency for DIS to evolve. Moreover, the equilibrium at F_{23} approaches the $ALLD$ corner as c/b decreases, which indicates that the evolution of DIS becomes easier on the $ALLD$-DIS edge as the cost to benefit ratio of cooperation decreases.
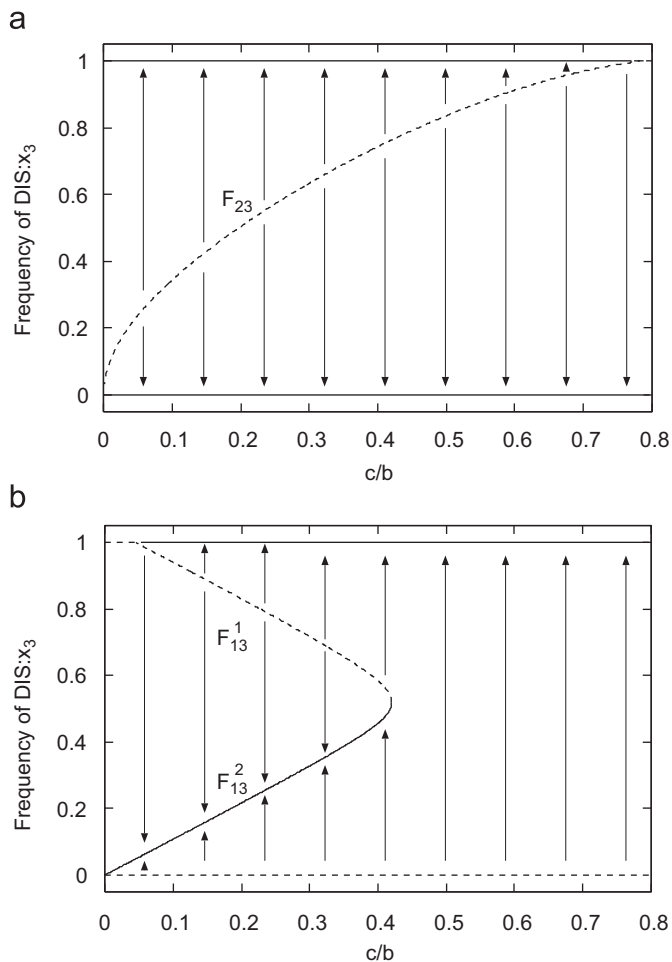
a

b

Fig. 2. (a) Bifurcation diagram of the equilibria on the *ALLD-DIS* edge: the solid line indicates a stable equilibrium and the dashed line indicates an unstable equilibrium ($\hat{\varepsilon} = 0.99$ and $w = 0.95$). (b) Bifurcation diagram of the equilibria on the *ALLC-DIS* edge: a solid line indicates a stable equilibrium and a dashed line indicates an unstable equilibrium ($\hat{\varepsilon} = 0.99$ and $w = 0.95$).

3.2. Evolutionary dynamics along the ALLC-DIS edge

Second, we investigate the evolutionary dynamics on the *ALLC-DIS* edge in the simplex S_3. Using numerical calculation, we illustrate the evolutionary dynamics on the edge in Fig. 2(b). The figure shows that, on the *ALLC-DIS* edge, an unstable internal equilibrium exists at F_{13}^1 and a stable internal equilibrium at F_{13}^2 in addition to the *ALLC* corner and the *DIS* corner. Put more precisely, if the cost-to-benefit ratio of cooperation is large (about $c/b > 0.4$), no internal equilibrium exists, and evolutionary dynamics always converge to the *DIS* corner. On the other hand, if the ratio is extremely small (about $c/b < 0.05$), the stable equilibrium at F_{13}^2 alone exists and then evolutionary dynamics always converge to F_{13}^2. Under the medium cost-to-benefit ratio of cooperation (about $0.05 < c/b < 0.4$), two internal equilibria exist. In this case, evolutionary dynamics converge to the *DIS* corner if there initially exist sufficiently many *DIS* strategists; otherwise, it converges to the stable internal equilibrium at F_{13}^2.

Furthermore, the figure shows that, as c/b decreases, F_{13}^1 approaches the *DIS* corner and F_{13}^2 approaches the *ALLC* corner, which indicates that the evolution of *DIS* becomes difficult as the cost-to-benefit ratio of cooperation decreases.

3.3. Global evolutionary dynamics

Next, we show the global dynamics of the frequency of the three strategies using numerical simulations.

In Fig. 3, we show, as a function of cost-to-benefit ratio of cooperation, c/b, the probability that evolutionary dynamics converges to each convergent point using 10 000 numerical simulation runs starting at random initial frequencies of the strategies. That probability is equivalent to the size of the basin of attraction for each convergent point. The figure shows that there exist three convergent points of the evolutionary dynamics: the *DIS* corner; the *ALLD* corner; and an equilibrium at F_{13}^2 where *DIS* and *ALLC* coexist.

Fig. 3 shows that, if $c/b > 0.8$ (region I in the figure), then all the dynamics converge to the *ALLD* corner. In this case, equilibria do not exist at F_{23}, F_{13}^1 or F_{13}^2 (see Fig. 2) and the *ALLD* corner is the only attractor of the evolutionary dynamics.

Evolutionary dynamics converge either to the *DIS* corner or to the *ALLD* corner depending on the initial frequency of the strategies if $0.4 < c/b < 0.8$ (region II in Fig. 3). The evolutionary dynamics for this case are illustrated in Fig. 4(a). In this case, an equilibrium exists at F_{23}, but no equilibria exist at F_{13}^1 or F_{13}^2 (see Fig. 2). Moreover, both convergence points, the *DIS* corner and
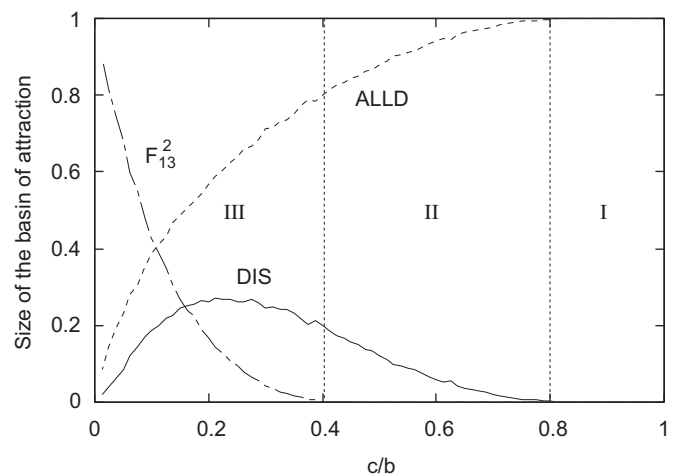

Fig. 3. The size of the basin of attraction for each of the three convergent points is plotted as a function of cost-to-benefit ratio of cooperation ($\hat{\varepsilon} = 0.99$ and $w = 0.95$). The solid line indicates the size of the basin of attraction for *DIS*; the dash-dotted line indicates that of attraction for F_{13}^2; and the dashed line indicates that of attraction for *ALLD*. Each line represents an average value over 10 000 simulation runs. Region I: No equilibrium exists at F_{23}, F_{13}^1 or F_{13}^2. Region II: An equilibrium exists at F_{23} and no equilibrium exists at F_{13}^1 or F_{13}^2. Region III: Equilibria exist at F_{23}, F_{13}^1 and F_{13}^2.
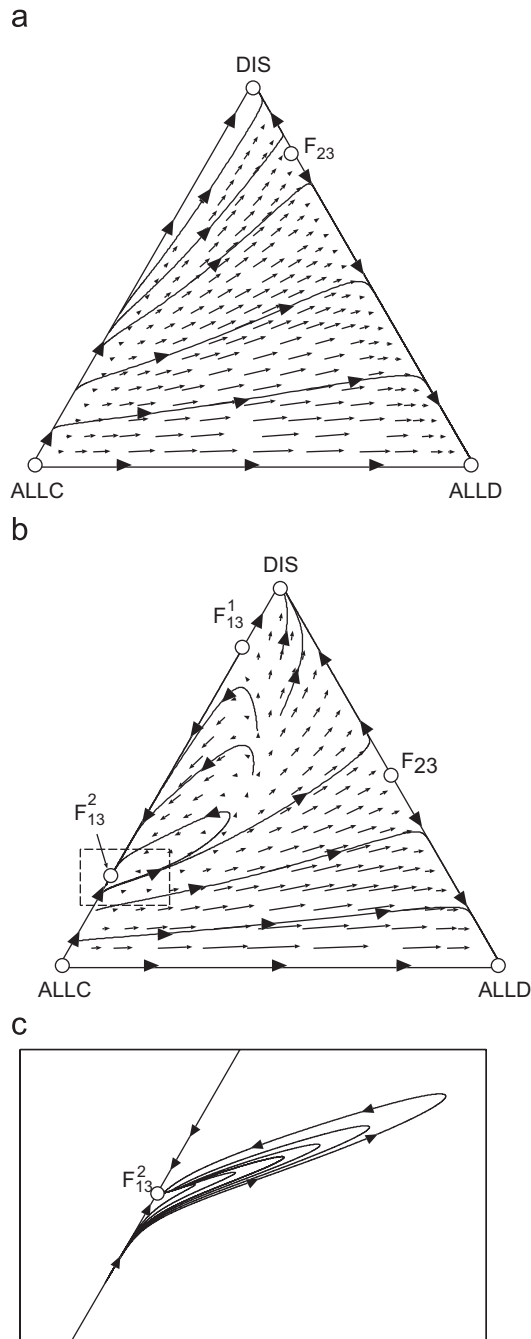
Fig. 4. Evolutionary dynamics of the frequency of the three strategies ($\hat{\varepsilon} = 0.99$ and $w = 0.95$): circles represent equilibria. (a) $c/b = 0.5$. (b) $c/b = 0.2$. (c) Evolutionary dynamics in the vicinity of the equilibrium at F_{13}^2 given in (b).

the *ALLD* corner, are asymptotically stable. The dynamics converge to the *DIS* corner when sufficiently numerous *DIS* exist initially, and to the *ALLD* corner otherwise. Furthermore, the size of the basin of attraction for *DIS* increases monotonically as c/b decreases.

The evolutionary dynamics converge to one of the three convergence points if $c/b < 0.4$ (region III in Fig. 3). The evolutionary dynamics for this case are illustrated in Fig. 4(b), which shows that the *DIS* corner and the *ALLD* corner are asymptotically stable. On the other hand, the

equilibrium at F_{13}^2 is not asymptotically stable[4]: However, even if perturbations occur, the dynamics at the equilibrium eventually revert to the equilibrium. For example, even if a perturbation at F_{13}^2 to produce ALLD and to decrease the frequency of DIS occurs, the newborn ALLD strategists will be driven out by DIS and thereby the coexistence of DIS and ALLC will be restored. In this sense, the equilibrium at F_{13}^2 is robust.

Compared with the dynamics under image scoring in two-person games (Panchanathan and Boyd, 2003), the dynamics in three-person games have the following two remarkable features. First, evolutionary dynamics can converge to the *DIS* corner in three-person games, although that never occurs in two-person games. This convergence results from the difference in evolutionary stability of *DIS* under image scoring between two-person and three-person games. As described previously, under image scoring, *DIS* can be an ESS in three-person games, whereas *DIS* cannot in two-person games (Panchanathan and Boyd, 2003). Second, in three-person games, the equilibrium at F_{13}^2 where *DIS* and *ALLC* coexist is robust in that, if perturbations to decrease the frequency of *DIS* occur, the dynamics at the equilibrium eventually revert to the equilibrium. On the other hand, in two-person games, the equilibrium at which *DIS* and *ALLC* coexist is not robust, and the dynamics at the equilibrium converge eventually to the *ALLD* corner if the perturbations occur (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2007; Brandt and Sigmund, 2006). That is, under image scoring, a society in which *DIS* and *ALLC* coexist can be maintained in three-person games, but cannot be maintained in two-person games. In other words, indirectly reciprocal cooperation can be formed and maintained, even under image scoring in three-person games, although it cannot be in two-person games.

4. Discussion

In this study, we have investigated the evolution of indirect reciprocity under image scoring in three-person games. We have particularly examined the case in the presence of error in implementation. Those analyses have revealed that indirect reciprocity can be formed and maintained under image scoring in three-person games, although results of previous works (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2007; Brandt and Sigmund, 2006) have shown that indirect reciprocity cannot in two-person games.

In two-person games, *DIS* is not stable against invasion by *ALLC* and coexists with *ALLC* as a stable polymorphism, but the polymorphism is vulnerable to perturbations (e.g. neutral drift or mutation) (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2007). In a population dominated by *DIS* strategists, a rare *ALLC* strategist achieves

---

[4]The equilibrium at F_{13}^2 is not asymptotically stable but Lyapunov stable because loops beneath the equilibrium become smaller as approaching the equilibrium (see Fig. 4(c)).

higher fitness because the invading *ALLC* strategist always intends to cooperate and thereby maintains a good reputation, never to be hurt by the incumbent *DIS* strategists except for erroneous defections, whereas the *DIS* strategists hurt each other in retaliation for others' erroneous defections. Therefore, initially a few *ALLC* strategists can increase their population in the *DIS* population. On the other hand, a few *DIS* strategists can invade the *ALLC* population because the invading *DIS* strategist can refuse to cooperate but still receive cooperation by the incumbent *ALLC* strategists. Consequently, the frequency of *DIS* and *ALLC* converges to a stable polymorphic equilibrium. However, if a non-adaptive process such as mutation or neutral drift were introduced, *ALLD* strategists would be able to invade the stable polymorphic equilibrium. Therefore, *ALLD* is the only long-term viable outcome.

In three-person games, as we have shown in this paper, phenomena different from those in two-person games are observed.

First, *DIS* is an ESS if the cost-to-benefit ratio of cooperation is in some intermediate range, i.e., *DIS* can resist the invasion by both *ALLD* and *ALLC*. The remarkable point is that *DIS* is stable against invasion by *ALLC*. In three-person games, a few invading *ALLC* strategists, who retain their good reputation, mostly belong to a group with two incumbent *DIS* strategists, who lose their good reputations because of retaliatory defections, in the population dominated by *DIS*. In this group, a *DIS* strategist defects in response to the bad reputation of the other *DIS* strategist even if the *ALLC* strategist has a good reputation. That is, a few invading *ALLC* strategists cannot avoid being caught in the retaliatory defection in three-person games. Therefore, the invading *ALLC* strategist cannot attain higher fitness than incumbent *DIS* strategists.

Second, in three-person games, there exists a polymorphic equilibrium at which *DIS* coexists with *ALLC*, as in two-person games. However, unlike two-person games, the polymorphic equilibrium is robust even if a non-adaptive process, such as mutation or neutral drift, is introduced. In terms of the evolution of cooperation, the polymorphic equilibrium is more important than the *DIS* equilibrium because, in the *DIS* equilibrium, the frequency of *DIS* with good reputation goes to zero over time: no *DIS* cooperates. Only the polymorphic equilibrium can attain high-level cooperation. (In Appendix B, we show the frequencies of cooperation on the simplex S_3 by density plot.)

At the polymorphic equilibrium, cooperation is maintained by an interesting role-sharing that occurs between *DIS* and *ALLC*. Specifically in this role-sharing, *ALLC* retains the good reputation of *DIS* so that cooperation continues. (Note that, without *ALLC*, *DIS* strategists hurt each other in response to erroneous defection; then all *DIS* strategists come to defect in the long run.) On the other hand, *DIS* provides protection against invasion by *ALLD*

in role-sharing. This role-sharing is reminiscent of some findings from the previous model of direct reciprocity that involves a so-called *nucleus-shield* (Lomborg, 1996). In our model, *DIS* is the shield that keeps out *ALLD*, while *ALLC* is the backbone of indirect reciprocal cooperation that maintains the good reputation of the population.

Many theoretical studies of two-person games (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2004, 2007; Brandt and Sigmund, 2004, 2006; Takahashi and Mashima, 2006) have concluded that indirectly reciprocal cooperation cannot be established under image scoring. They have stated that people should use more complicated reputation criteria (e.g. standing (Sugden, 1986)) to establish cooperation. However, which reputation criterion people actually use, image scoring or more complicated one, remains a controversial issue in experimental studies (Milinski et al., 2001; Bolton et al., 2005). Regarding two-person games, some experimental studies of human subjects (Wedekind and Milinski, 2000; Milinski et al., 2001) have demonstrated that image-scoring is widely used. Nonetheless, other experimental studies have shown that human beings use more complicated reputation criteria than image scoring (Bolton et al., 2005). Based on the results of this study, we claim that indirect reciprocal cooperation can be established even under image scoring when three individuals interact in a single group. This conclusion might suggest the possibility that experimental studies of indirect reciprocity in three-person games find a very different result from that in two-person games.

Acknowledgments

Appendix A

Here we derive the fitness for each strategy, which is defined as the individual's expected total payoff during a generation. The payoff for an individual in a round is determined by the probability that the focal individual cooperates and by the probability that an opponent of the focal individual cooperates (see Eqs. (1) and (2)).

Because *ALLD* never cooperates and *ALLC* always intends to cooperate, the respective probabilities that *ALLD* and *ALLC* cooperate are 0 and $\hat{\varepsilon}$. Furthermore, *DIS* intends to cooperate only when the *two* opponents have reputation G. Let $g(t)$ be the frequency of individuals with reputation G among the whole population at round t. Then, the probability that *DIS* cooperates at round t is $\hat{\varepsilon}g(t)^2$. Moreover, we represent the frequency of individuals with reputation G at round t among *ALLC*, *ALLD* and *DIS* strategies, respectively, as $g_1(t)$, $g_2(t)$ and $g_3(t)$. In this

case, $g(t) = g_1(t)x_1 + g_2(t)x_2 + g_3(t)x_3$. At the first round, all individuals have reputation G; therefore, $g(1) = g_1(1) = g_2(1) = g_3(1) = 1$. Because $ALLC$ always intends to cooperate, $g_1(t) = \hat{\varepsilon}$ for $t \geqslant 2$; because $ALLD$ always defects, $g_2(t) = 0$ for $t \geqslant 2$. Furthermore, because DIS intends to cooperate at round $t-1$ with the probability $\hat{\varepsilon}g(t-1)^2$, $g_3(t) = \hat{\varepsilon}g(t-1)^2$ for $t \geqslant 2$, and so $g_3(2) = \hat{\varepsilon}$ and $\lim_{t \to \infty} g_3(t) = (1 - 2\hat{\varepsilon}^2 x_1 x_3 - \sqrt{1 - 4\hat{\varepsilon}^2 x_1 x_3})/2\hat{\varepsilon}x_3^2$, i.e., the frequency of individuals with reputation G among DIS approaches the above value over time.

On the other hand, an opponent of the focal individual intends to cooperate only in the following two situations: (1) the opponent has the $ALLC$ strategy, the probability of which is x_1; and (2) the opponent has the DIS strategy, and both the focal individual and the other opponent have reputation G, the probability of which is $g_i(t)g(t)x_3$, where $i \in \{1, 2, 3\}$.

Therefore, the expected payoff at round t for the three strategies, $ALLC$, $ALLD$ and DIS, represented respectively as $f_1(t)$, $f_2(t)$ and $f_3(t)$, are

$$f_1(t) = \hat{\varepsilon}b[x_1 + g_1(t)g(t)x_3] - \hat{\varepsilon}c, \tag{A.1}$$

$$f_2(t) = \hat{\varepsilon}b[x_1 + g_2(t)g(t)x_3], \tag{A.2}$$

$$f_3(t) = \hat{\varepsilon}b[x_1 + g_3(t)g(t)x_3] - \hat{\varepsilon}cg(t)^2. \tag{A.3}$$

The fitness for strategy i ($i = 1, 2, 3$), which is defined as its expected total payoff during a generation, is

$$f_i = \sum_{t=1}^{\infty} w^{t-1} f_i(t). \tag{A.4}$$

We cannot get the exact calculation of f_i. Therefore, we obtain those values approximately by numerical calculation $f_i = \sum_{t=1}^{T} w^{t-1} f_i(t)$ with finite T. This approximation

does not change the results essentially as far as T is sufficiently large, because $0 \ll w < 1$ and $f_i(t)$ is bounded. Throughout this study, we set $T = 10\,000$.

Appendix B

We show the density plot of the frequency of cooperation on the simplex S_3 in Fig. 5. As the figure shows, the frequency of cooperation at the DIS equilibrium is very low, but the frequency at the polymorphic equilibrium where DIS coexists with $ALLC$ is very high.

References

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York.

Bolton, G.E., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. J. Public. Econ. 89, 1457–1468.

Boyd, R., Richerson, P.J., 1988. The evolution of reciprocity in sizable groups. J. Theor. Biol. 132, 337–356.

Brandt, H., Sigmund, K., 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theor. Biol. 231, 475–486.

Brandt, H., Sigmund, K., 2006. The good, the bad and the discriminator—errors in direct and indirect reciprocity. J. Theor. Biol. 239, 183–194.

Bshary, R., 2002. Biting cleaner fish use altruism to deceive image scoring clients. Proc. R. Soc. Lond. Ser. B 269, 2087–2093.

Bshary, R., Grutter, A.S., 2006. Image scoring and cooperation in a cleaner fish mutualism. Nature 441, 975–978.

Dugatkin, L.A., 1990. N-person games and the evolution of co-operation: a model based on predator inspection in fish. J. Theor. Biol. 142, 123–135.

Eriksson, A., Lindgren, K., 2005. Cooperation driven by mutations in multi-person prisoner's dilemma. J. Theor. Biol. 232, 399–409.

Fishman, M.A., 2003. Indirect reciprocity among imperfect individuals. J. Theor. Biol. 225, 285–292.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Hauert, C., Monte, S.D., Hofbauer, J., Sigmund, K., 2002. Volunteering as red queen mechanism for cooperation in public goods games. Science 296, 1129–1132.

Hauert, C., Holmes, M., Doebeli, M., 2006. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. Proc. R. Soc. Lond. Ser. B 273, 2565–2570.

Joshi, N.V., 1987. Evolution of cooperation by reciprocation within structured demes. J. Genet. 66, 69–84.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond. Ser. B 268, 745–753.

Lindgren, K., Johansson, J., 2001. Coevolution of strategies in n-person prisoner's dilemma. In: Crutchfield, J.P., Schuster, P. (Eds.), Evolutionary Dynamics: Exploring the Interplay of Selection, Accident, Neutrality, and Function. Santa Fe Institute Studies in the Sciences of Complexity Series.

Lomborg, B., 1996. Nucleus and shield: the evolution of social structure in the iterated prisoner's dilemma. Amer. Sociol. Rev. 61, 278–307.

Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.J., 2001. Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. Lond. Ser. B 268, 2495–2501.

Nowak, M.A., Sigmund, K., 1998a. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.

Nowak, M.A., Sigmund, K., 1998b. The dynamics of indirect reciprocity. J. Theor. Biol. 194, 561–574.

Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.

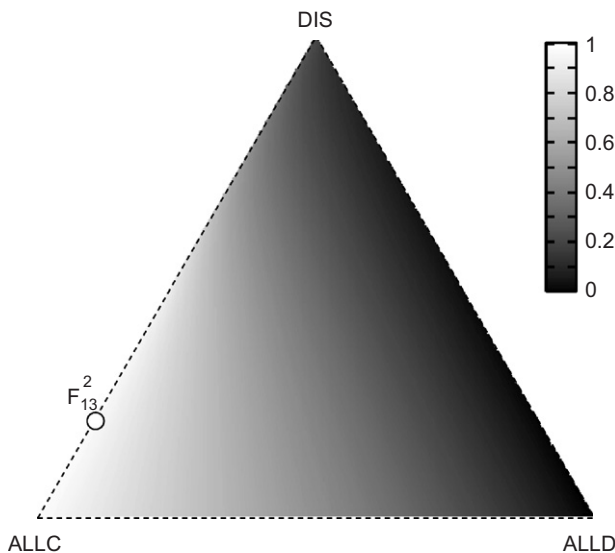Fig. 5. Average frequency of cooperation in a generation in the simplex S_3 ($\hat{\varepsilon} = 0.99$ and $w = 0.95$): the frequency is illustrated as gray scale (light is high and dark is low) for each point in the simplex. The point at F_{13}^2 in this figure indicates the equilibrium at F_{13}^2 in Fig. 4 (b).

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. J. Theor. Biol. 231, 107–120.

Ohtsuki, H., Iwasa, Y., 2007. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. J. Theor. Biol., in press.

Ostrom, E., Burger, J., Field, C.B., Norgaard, R.B., Policansky, D., 1999. Sustainability—revisiting the commons: local lessons, global challenges. Science 284, 278–282.

Panchanathan, K., Boyd, R., 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J. Theor. Biol. 224, 115–126.

Sugden, R., 1986. The Evolution of Rights Co-operation and Welfare. Blackwell, Oxford.

Suzuki, S., Akiyama, E., 2007. Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. J. Theor. Biol. 245, 539–552.

Suzuki, S., Akiyama, E., submitted for publication. Evolutionary stability of first-order-information indirect reciprocity in sizable groups.

Takahashi, N., Mashima, R., 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. J. Theor. Biol. 243, 418–436.

Taylor, P., Jonker, L., 1978. Evolutionary stable strategies and game dynamics. Math. Biosci. 40, 145–156.

Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.