PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2007) **274**, 689–695
doi:10.1098/rspb.2006.3759
Published online 12 December 2006

Tag-based indirect reciprocity by incomplete social information

Naoki Masuda[1,*] **and Hisashi Ohtsuki**[2]

[1]*Amari Research Unit, RIKEN Brain Science Institute, 2-1, Hirosawa, Wako, Saitama 351-0198, Japan*
[2]*Department of Biology, Faculty of Sciences, Kyushu University, Hakozaki 6-10-1, Fukuoka 812-8581, Japan*

Evolution of altruistic behaviour in interacting individuals is accounted for by, for example, kin selection, direct reciprocity, spatially limited interaction and indirect reciprocity. Real social agents, particularly humans, often take actions based on similarity between themselves and others. Although tag-based indirect reciprocity in which altruism occurs exclusively among similar flocks is a natural expectation, its mechanism has not really been established. We propose a model of tag-based indirect reciprocity by assuming that each player may note strategies of others. We show that tag-based altruism can evolve to eradicate other strategies, including unconditional defectors for various initial strategy configurations and parameter sets. A prerequisite for altruism is that the strategy is sometimes, but not always, visible to others. Without visibility of strategies, policing does not take place and defection is optimal. With perfect visibility, what a player does is always witnessed by others and cooperation is optimal. In the intermediate regime, discriminators based on tag proximity, rather than mixture of generous players and defectors, are most likely to evolve. In this situation, altruism is realized based on homophily in which players are exclusively good to similar others.

Keywords: altruism; evolutionary game; indirect reciprocity; homophily

1. INTRODUCTION

According to classical game theory, defection rather than cooperation is the optimal behaviour from individualistic points of view. However, altruistic behaviour evolves in real social systems, particularly in human societies (Fehr & Fischbacher 2003). Altruism is explained by mechanisms such as kin selection (Hamilton 1964), assortative mating (Eshel & Cavalli-Sforza 1982; Bergstrom 2003), direct reciprocity (Trivers 1971; Axelrod & Hamilton 1981; Axelrod 1984), spatial reciprocity based on short-range interaction (Nowak & May 1992) and indirect reciprocity (Nowak & Sigmund 1998*a*,*b*, 2005; Brandt & Sigmund 2004, 2005; Ohtsuki & Iwasa 2004). In particular, indirect reciprocity occurs if each player carries an image score that indicates how often the player has cooperated before. Then, unconditional cooperators with high image scores are trusted by others, but they are too good to be resistant to invasion by unconditional defectors. Defectors have low scores and cannot count on altruism from others. Discriminators with moderate scores can evolve because good players trust them and defectors cannot exploit them.

In many situations, individuals are equipped with biological and social tags according to which they are categorized. Similar birds tend to flock together based on tags, which is assortative interaction (McPherson & Smith-Lovin 1987; Marsden 1988; McPherson *et al.* 2001; Newman 2003). Intuitively, players may exclusively help similar others to lead to indirect reciprocity. Imagine that we meet a stranger *P* from the same town or school by chance. Then, we may be inclined to help *P* even if we do not expect to see *P* later on. However, defection is theoretically better in such a one-shot encounter.

Tag-based reciprocity was proposed as a mechanism of indirect reciprocity (Riolo *et al.* 2001). Each player has a tolerance value as strategy. Altruistic behaviour happens when the difference between the opponent's tag and the tag of the focal player does not exceed the tolerance of the focal player. Altruism can evolve among players with close tags even if they have not met before. A key factor for tag-based altruism in this model is the assumption that players with the same tag value always help each other. If this condition is removed, which may be more realistic, cooperation does not evolve (Roberts & Sherratt 2002).

Nevertheless, these debates do not exclude the possible importance of tag-based altruism. In combination with population viscosity, which is known to elicit altruism in the Prisoner's Dilemma (Nowak & May 1992), use of tags reinforces altruism (Hammond & Axelrod 2006). Another scenario that revives tag-based altruism is that the tag and the tolerance genes sit on different alleles and inheritance of these genotypes are only loosely coupled. Thus, if many tag values are allowed and a population is spatially organized, tag-based altruism is viable (Jansen & van Baalen 2006).

In this work, we propose a mechanism of indirect reciprocity based on tags, without resorting to a prefabricated cooperative tendency (Riolo *et al.* 2001) and spatial structure of populations (Nowak & May 1992; Hammond & Axelrod 2006; Jansen & van Baalen 2006). The main assumption is that players may observe the tolerance (strategy) of opponents. In reality, third parties may communicate the information about others' strategies to a focal player. We show that tag-based indirect reciprocity emerges for a variety of initial configurations when tolerances are observable with moderate probabilities.

* Author and address for correspondence: Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan. (masuda@mist.i.u-tokyo.ac.jp).

Received 18 August 2006
Accepted 22 October 2006

2. MODEL: STRATEGY AND PAY-OFF

We assume n individuals involved in the evolutionary Prisoner's Dilemma game. In each round, a donor and a recipient are taken from the population randomly. If the donor decides to donate, the donor loses cost c and the recipient gains benefit b, which is larger than c. If the donor refuses to donate, both the donor and the recipient get nothing. A generation consists of a sufficiently small number of rounds that prohibits two specific players from interacting more than once on average. Therefore, there is no room for direct reciprocity. After a generation, the strategy of players with larger total pay-offs is more likely to be disseminated, whereas poor strategies are eliminated. Since $b>c$, the population pay-off is maximized when everybody is always altruistic. However, unconditional defection, but not cooperation, is the evolutionary stable strategy (ESS) (Axelrod & Hamilton 1981; Axelrod 1984); here is a dilemma.

We assume that the ith individual ($1\leq i\leq n$) has tag w_i and tolerance μ_i. Everybody can see others' tags. Tags may take nominal or graduated values. In either case, the distance between two tags $0\leq l(w_i,w_j)\leq l_{\max}$ is defined, where $l_{\max}$ is the maximal distance. The tag w_i can be assumed to vary in time (Riolo *et al.* 2001; Jansen & van Baalen 2006), which is typically the case in, for example, the preference of political parties in democratic regimes and fashions in clothing. Tags evolve much more slowly than do tolerances in other situations. For example, inborn biological or social traits and strong beliefs can be considered invariant for a player, whereas tolerances can adapt via social learning on a relatively short time-scale. We focus on the latter case; we consider the evolution of tolerances with tags fixed.

Suppose that player i and player j are, respectively, chosen as donor and recipient in a single round. Then, the player i may donate to the player j if i and j are close enough according to i's criterion. In the previous model (Riolo *et al.* 2001), i pays c to benefit j by b if $l(w_i,w_j)\leq\mu_i$, where μ_i is the tolerance level of i. Players with $\mu_i\geq l_{\max}$ always donate, whereas players with $\mu_i<0$ never donate. These players do not resort to tags. Riolo *et al.* (2001) excluded unconditional defectors by setting $\mu_i\geq 0$ and observed indirect reciprocity. However, we allow negative μ_i to investigate whether tag users can evolve in the presence of unconditional defectors ($\mu_i<0$). Without additional mechanisms, unconditional defectors outperform unconditional cooperators and tag users; hence, there is no altruism (Roberts & Sherratt 2002).

To explore the possibility of tag-based indirect reciprocity, we introduce the probability q that the donor i detects μ_j, namely, the tolerance of the recipient j. We can tune the reputations of others' tolerances to be public with a large q or private with a small q (Nowak & Sigmund 1998*a*,*b*; Brandt & Sigmund 2004, 2005), whereas perfect social information ($q=1$) is assumed in other models (e.g. Ohtsuki & Iwasa 2004). If i knows that $l(w_i,w_j)>\mu_j$, i is not motivated to be nice to j. Note that this is not because i does not expect cooperation back from j; i and j hardly interact more than once per generation. Rather, this is because i expects that j will not be nice to other players whose tags are close to w_i. If other players do the same, natural selection will eliminate μ_j, which may lead to an altruistic population.

If $q=0$, the donor never sees the tolerance of the recipient. In this case, $\mu<0$ will be stable (Roberts & Sherratt 2002). If $q=1$, the tolerance of everybody is transparent to the entire population. Thus, small μ_i will disappear rapidly, and everybody will be eventually altruistic ($\mu=l_{\max}$). We examine what occurs for intermediate values of q.

Table 1. Probability that two players meet in the two-tag model.

	tag of the opponent	
tag of the focal player	w^a	w^b
w^a	$t+h(1-t)$	$(1-h)(1-t)$
w^b	$(1-h)t$	$1-t+ht$

3. TWO-TAG MODEL

We begin with a minimal model in which each player has tag w^a or w^b. Such a minimal model was used to analyse the replicator dynamics of the original model by Riolo *et al.* (Traulsen & Schuster 2003). We define $l(w^a,w^a)=l(w^b,w^b)=0$ and $l(w^a,w^b)=1$. It suffices to consider three tolerance values $\mu=-1$, 0 and 1. Players with $\mu=-1$ are unconditional defectors. Players with $\mu=0$ are tag users. Players with $\mu=1$ are unconditional cooperators if $q=0$. When tolerances of opponents are visible ($q>0$), these players do not necessarily cooperate. We call them generous players. We could introduce unconditional cooperators and players who neglect others' tolerances even when available. However, these strategies are outperformed by players using tags and tolerances. Therefore, we consider only the three strategies ($\mu=1$, 0 and -1).

Suppose that a population consists of n players and that nt ($n(1-t)$) players are endowed with tag w^a (w^b). We denote by p^a_μ the probability that a player with tag w^a, who we call w^a-player, has tolerance μ. Similarly, a w^b-player has tolerance μ with probability p^b_μ. If the population is well mixed, any player meets a w^a-player (w^b-player) and tolerance μ with probability $tp^a_\mu((1-t)p^b_\mu)$. Since the tag is visible to others, players may prefer to interact with conspecifics. Such assortative interaction may stem from spatial structure of the population. To take assortativity into account, we assume that the two players with the same tag can meet more frequently. Precisely, a player who meets an opponent with the opposite tag looks for an alternative opponent with the same tag with probability h ($0\leq h\leq 1$). Thus, a w^a-player interacts with a w^a-player (w^b-player) with probability $t+h(1-t)$ ($(1-h)(1-t)$), whereas a w^b-player interacts with a w^a-player (w^b-player) with probability $(1-h)t$ ($1-t+ht$) (see table 1; Eshel & Cavalli-Sforza 1982; Bergstrom 2003). Non-assortative interaction is reproduced by $h=0$, and $h=1$ defines completely assortative interaction in which players with opposite tags never interact.

Disregarding normalization, the pay-off $\Pi^a_\mu(\Pi^b_\mu)$ for a w^a-player (w^b-player) with tolerance μ is given by

$$\Pi^a_1=(b-cq)\left[(t+h-ht)\left(p^a_0+p^a_1\right)+(1-h)(1-t)p^b_1\right]-c(1-q), \tag{3.1}$$

$$\Pi^a_0=b\left[(t+h-ht)(p^a_0+p^a_1)+(1-q)(1-h)(1-t)p^b_1\right]-c(1-q)(t+h-ht)-cq(t+h-ht)\left(p^a_0+p^a_1\right), \tag{3.2}$$

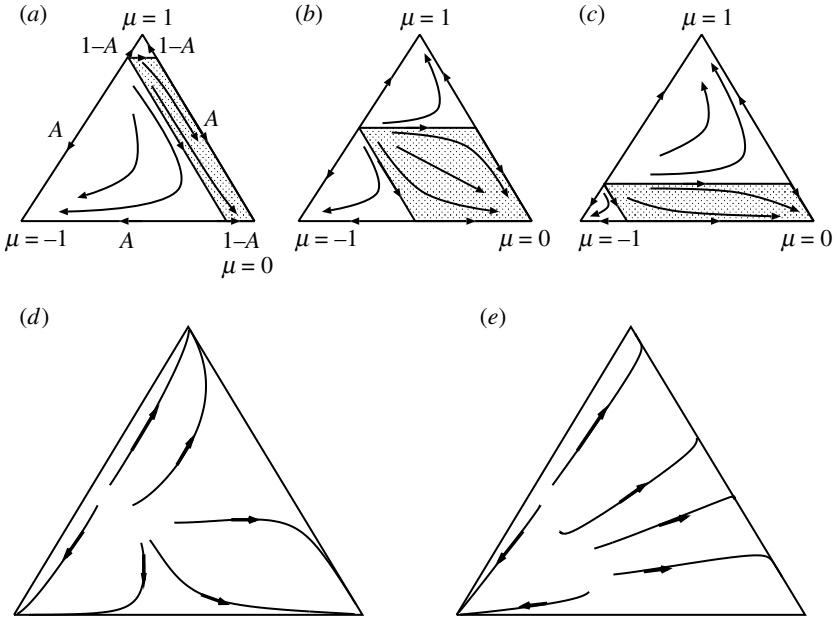

Figure 1. Replicator dynamics with binary tags. (*a–c*) Schematics of population dynamics for $h=0$ and (*a*) small, (*b*) intermediate, and (*c*) large q. (*d*, *e*) Comparison of different degrees of assortativity with $b=1.0$, $c=0.3$ and $q=0.5$. (*d*) $h=0$ and (*e*) $h=0.8$.

$$\Pi^a_{-1} = b(1-q)\left[(t+h-ht)\left(p^a_0+p^a_1\right)+(1-h)(1-t)p^b_1\right], \tag{3.3}$$

$$\Pi^b_1 = (b-cq)\left[(1-t+ht)\left(p^b_0+p^b_1\right)+(1-h)tp^a_1\right]-c(1-q), \tag{3.4}$$

$$\begin{aligned}\Pi^b_0 = {} & b\left[(1-t+ht)\left(p^b_0+p^b_1\right)+(1-q)(1-h)tp^a_1\right] \\ & -c(1-q)(1-t+ht)-cq(1-t+ht)\left(p^b_0+p^b_1\right)\end{aligned} \tag{3.5}$$

and

$$\Pi^b_{-1} = b(1-q)\left[(1-t+ht)\left(p^b_0+p^b_1\right)+(1-h)tp^a_1\right]. \tag{3.6}$$

Then, we examine the replicator dynamics (Taylor & Jonker 1978; Hofbauer & Sigmund 1998) of the population density described by

$$\dot{p}^a_\mu = p^a_\mu\left(\Pi^a_\mu - \sum_{\mu'=-1,0,1} p^a_{\mu'}\Pi^a_{\mu'}\right), \tag{3.7}$$

$$\dot{p}^b_\mu = p^b_\mu\left(\Pi^b_\mu - \sum_{\mu'=-1,0,1} p^b_{\mu'}\Pi^b_{\mu'}\right). \tag{3.8}$$

The pay-off of w^a-players depends on the tolerance distribution of w^b-players, who do not compete with w^a-players, and vice versa. There are six variables (p^a_{-1}, p^a_0, p^a_1, p^b_{-1}, p^b_0 and p^b_1). Since we are interested in competition among tolerances with tags fixed, the number of w^a-players ($=nt$) and that of w^b-players ($=n(1-t)$) is preserved. Then, we obtain $p^a_{-1}+p^a_0+p^a_1=1$, $p^b_{-1}+p^b_0+p^b_1=1$, and the replicator system represented by equations (3.7) and (3.8) is four-dimensional.

(a) *Symmetric strategies*

Let us start with a simple case in which w^a-players equal w^b-players in number ($t=0.5$) and the initial condition is symmetric: $p^a_\mu = p^b_\mu$ ($\mu=-1$, 0 and 1). Then, the dynamics represented by equations (3.7) and (3.8) preserve the relations $p_\mu \equiv p^a_\mu = p^b_\mu$ and $\Pi_\mu \equiv \Pi^a_\mu = \Pi^b_\mu$ ($\mu=-1, 0$ and 1), and are equivalent to the two-dimensional replicator dynamics with p_{-1}, p_0 and p_1 ($p_{-1}+p_0+p_1=1$). Equations (3.1)–(3.6) are reduced to

$$\Pi_1 = (b-cq)\left(\frac{1+h}{2}p_0+p_1\right)-c(1-q), \tag{3.9}$$

$$\Pi_0 = \left[b(1-q)+\frac{(b-c)q(1+h)}{2}\right]p_1 + \frac{(b-cq)(1+h)}{2}p_0 - \frac{c(1-q)(1+h)}{2}, \tag{3.10}$$

and

$$\Pi_{-1} = b(1-q)\left(\frac{1+h}{2}p_0+p_1\right). \tag{3.11}$$

Then, we have

$$\Pi_1 > \Pi_0 \leftrightarrow (b-c)qp_1 > c(1-q), \tag{3.12}$$

$$\Pi_0 > \Pi_{-1} \leftrightarrow (b-c)q(p_0+p_1) > c(1-q), \tag{3.13}$$

and

$$\Pi_1 > \Pi_{-1} \leftrightarrow (b-c)q\left(\frac{1+h}{2}p_0+p_1\right) > c(1-q). \tag{3.14}$$

As schematically shown in figure 1, the tolerance distribution of a population is mapped to a point in the simplex defined by $p_0+p_1+p_2=1$ and $p_0, p_1, p_2 \geq 0$. On the line $p_{-1}=0$ (or $p_0+p_1=1$), we have $\Pi_1 > \Pi_0 \leftrightarrow p_1 > A$, where $A=c(1-q)/[(b-c)q] \geq 0$. Similarly, we derive $\Pi_1 > \Pi_{-1} \leftrightarrow p_1 > A$ on $p_0=0$ and $\Pi_0 > \Pi_{-1} \leftrightarrow p_0 > A$ on $p_1=0$. Therefore, if $A<1$, then $(p_{-1}, p_0, p_1)=(1,0,0),(0,1,0),(0,0,1)$ are ESSs. The condition $A<1$ is equivalent to $bq>c$, which coincides with the requirement

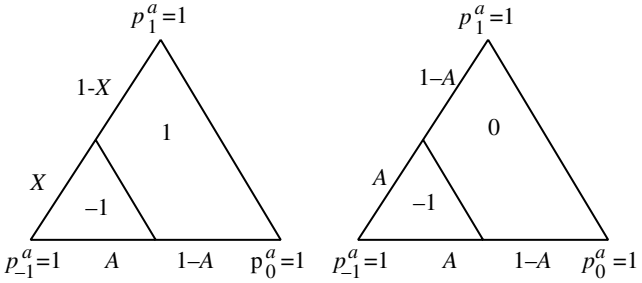

Figure 2. The best response for w^a-players when (*a*) $p_1^b > A$ and (*b*) $p_1^b < A$. We set $X = \left[A-(1-(t+h-ht))p_1^b\right]/(t+h-ht)$.

for altruism in the image scoring model (Nowak & Sigmund 1998*a*,*b*). If $bq \le c$, Π_{-1} is dominant, and unconditional defectors eventually occupy the population.

Let us concentrate on the case $bq > c$ ($A < 1$). In addition to the three corners of the simplex, $(p_{-1}, p_0, p_1) = (1-A, 0, A)$ is a degenerated source (see Appendix A). Dynamics with non-assortative interaction ($h = 0$) are compared in figure 1*a–c* for different values of q. For a small q ($bq \cong c$), most initial configuration leads to domination by unconditional defectors ($p_{-1} = 1$ and $p_0 = p_1 = 0$), as shown in figure 1*a*. Since players can keep their tolerances secret, it is better to behave selfishly (Roberts & Sherratt 2002). As q increases (equivalently, A decreases), tolerances are observed by others, and it is rational to be nice. An increase in q, indeed, enlarges the attractive basins of the unanimity of tag users ($p_0 = 1$ and $p_{-1} = p_1 = 0$) and that of generous players ($p_1 = 1$ and $p_{-1} = p_0 = 0$). By contrast, the attractive basin of $p_{-1} = 1$ shrinks. For intermediate q, the attractive basin of $p_0 = 1$ (shaded areas in figure 1*a–c*) is large. Particularly, when $q = 2c/(b+c)$ (namely, $A = 0.5$), the attractive basin of $p_0 = 1$ is largest and twice as large as those of $p_{-1} = 1$ and $p_1 = 1$ (figure 1*b*). When q is close to $2c/(b+c)$, tag users are likely to outperform generous players and defectors to establish tag-based altruism for many initial conditions. If q is even larger, being generous is optimal for many configurations (figure 1*c*).

Numerically calculated replicator dynamics for $q = 0.5$ are shown in figure 1*d*,*e* for two values of h. The results for $h = 0$ (non-assortative encountering) resemble figure 1*b*. When h is large (assortative encountering), $\mu = 0$ and 1 are not so distinct because players do not often play with others with the opposite tag. Then, the trajectories inside the simplex are more or less symmetric about $p_0 = p_1$ (figure 1*e*). By contrast, if players with the opposite tags are more likely to interact, $\mu = -1$ and 0 are less distinguished. Note that assortativity ($h > 0$) is not necessary for tag-based reciprocity.

(b) *Asymmetric strategies*

The assumption that w^a- and w^b-subpopulations share the identical number of players and the identical tolerance distributions is too restrictive. Here, we relax $t = 0.5$ and $p_i^a = p_i^b$ ($i = 0,1,2$). Now the population dynamics described by equations (3.7) and (3.8) are four dimensional. Since analytical treatment of the four-dimensional system is intractable, we look for pure-strategy ESSs. If a pure strategy of the w^a-subpopulation is the best response to the w^b-subpopulation with a pure strategy and vice versa, this pair of tolerances is an ESS. Therefore, we investigate compatibility of $3 \times 3 = 9$ pure strategies.

Let us fix the tolerance distribution of the w^b-players. Then, the best response for w^a-players can be derived from equations (3.1)–(3.3). The best response depending on the tolerance density of w^a-players is shown in figure 2*a* for $p_1^b > A$. Since $X \equiv \left[A-(1-(t+h-ht))p_1^b\right]/(t+h-ht) < A$ for $p_1^b > A$, the region defined by $p_1^a > A$ belongs to the basin of $p_1^a = 1$. The same conclusion obeys with a and b swapped and $t+h-ht$ replaced by $1-t+ht$. Starting from $p_1^a > A$ and $p_1^b > A$, generous players will eventually eradicate the other two strategies in both subpopulations. This extends the result for the symmetric case.

The best response for w^a-players when $p_1^b < A$ is shown in figure 2*b*. Since $p_1^b < A$, possible pure strategies for the w^b-subpopulation are $\mu = -1$ and 0. Figure 2*b* implies that the pure strategy $p_1^a = 1$ is incompatible with $p_{-1}^b = 1$ and $p_0^b = 1$. Similarly, $p_1^b = 1$ is incompatible with $p_{-1}^a = 1$ and $p_0^a = 1$. Therefore, we exclude $(p_{-1}^a, p_0^a, p_1^a, p_{-1}^b, p_0^b, p_1^b) = (1,0,0,0,0,1)$, $(0,1,0,0,0,1)$, $(0,0,1,1,0,0)$ and $(0,0,1,0,1,0)$. By contrast, four pure strategies: $(p_{-1}^a, p_0^a, p_1^a, p_{-1}^b, p_0^b, p_1^b) = S_1 : (1,0,0,1,0,0)$, $S_2 : (1,0,0,0,1,0)$, $S_3 : (0,1,0,1,0,0)$ and $S_4 : (0,1,0,0,1,0)$ are self-consistent stable steady states. Solutions S_1 (all defectors) and S_4 (all tag users) are symmetric, as discovered in §3*a*. With S_2 (w^a-defectors and w^b-tag users) or S_3 (w^a-tag users and w^b-defectors), players do not cooperate with dissimilar others. Consequently, how nice a w^a- (w^b-) player is to other w^a- (w^b-) players does not affect the pay-off and the tolerances of w^b- (w^a-) players. Since two groups are independent of each other, asymmetric solutions are allowed.

To investigate which of the five ESSs identified above is more feasible, we perform 3000 runs of evolutionary dynamics with different initial conditions. The initial tolerance density is chosen according to the uniform distribution on the simplex: $p_{-1}^x + p_0^x + p_1^x = 1$ ($p_{-1}^x, p_0^x, p_1^x \ge 0$) independently for $x = a$ and b, respectively. The initial tolerance of each player is picked according to the tolerance distributions generated in this way. We set $n = 300$, $b = 1.0$ and $c = 0.3$. One generation consists of $8n$ rounds so that each player appears eight times as donor and eight times as recipient on average. Since $8 \times 2 \ll n$, a specific pair of players hardly meet more than once, and direct reciprocity is ruled out. We study asynchronous updating: after each generation, we randomly choose a small number of players (equal to 8). For each of these players, the tolerance is replaced by that of player i with probability $\left(\pi_i - \min_{j=1}^n \pi_j\right) / \sum_{i=1}^n \left(\pi_i - \min_{j=1}^n \pi_j\right)$, where π_i ($1 \le i \le n$) is the ith player's generation pay-off. Players with larger π_i are more likely to bear offspring. The players with the minimal pay-off ($= \min_{j=1}^n \pi_j$) do not reproduce themselves.

The final proportion of each pure strategy is shown in figure 3. For small q, $\mu = -1$ results from most initial

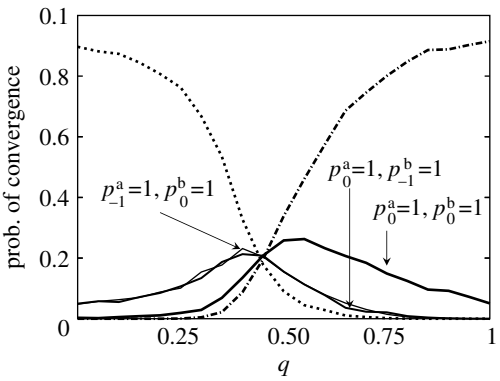

Figure 3. The size of the basin of attraction of five main pure strategies.

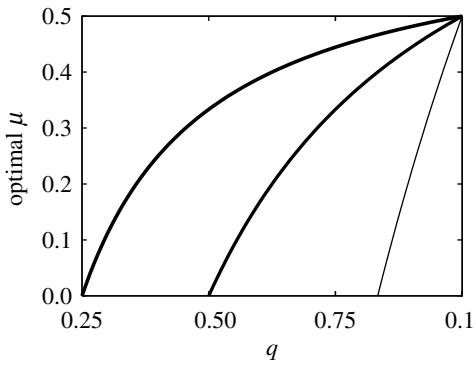

Figure 4. The best response in the continuous-tag model when tags and tolerances are uniformly distributed. $b/c=1.2$ (thinnest line), 2 (moderate line) and 4 (thickest line).

conditions for both subpopulations (dotted line). By contrast, $\mu=1$ is reached for both subpopulations when q is large (dash-dotted line). For intermediate q, three pure strategies comprising tag users can evolve. The probability that any of the three ESSs, with tag users fixates is pretty large (≥ 0.5) for $0.4 \leq q \leq 0.55$. This range of q agrees with $q=2c/(b+c) \cong 0.46$, which is the optimal q for prosperity of tag users in the symmetric population (§3*a*). The probability that any of the other four pure strategies dominates the population is tiny and is thus omitted in figure 3.

4. CONTINUOUS-TAG MODEL

Let us consider the case of continuously distributed tags, which is more realistic in some practical situations. Similar to Riolo *et al.* (2001), we assume that tags are uniformly distributed on [0,1], where 0 and 1 are identified, and that the interaction is non-assortative. The difference between two tags is naturally defined as the distance on the ring. Formally, $l(w_i,w_j)=\min(|w_i-w_j|,\ 1-|w_i-w_j|)$ and $0 \leq l(w_i,w_j) \leq l_{\max} \equiv 0.5$. We set $\mu_i \in [\mu_{\min}, \mu_{\max}]$, where $\mu_{\min}=-10^{-6}$ and $\mu_{\max}=0.5$. Players with negative μ_i are unconditional defectors. Along the lines of the two-tag model, it is straightforward to show that a population of players with $\mu_i=\mu$ $(1 \leq i \leq n)$ for any μ is not invaded by a player with a different tolerance, given $bq>c$. Such a unanimous population is an ESS for any μ. We are interested in which μ will evolve.

To see this, imagine that each player initially owns a tolerance value taken from the uniform distribution. Without losing generality, we assume that a reference player P_1 has tag $w=0.5$ and tolerance μ. Player P_1 potentially donates to others whose tag $w' \in [0.5-\mu,\ 0.5+\mu]$. Owing to the uniform tag density, a proportion of $2\Delta l$ players, or players with tag $w' \in [0.5+l,\ 0.5+l+\Delta l]$ or $w' \in [0.5-l-\Delta l,\ 0.5-l]$, are in the range of distance $[l,l+\Delta l]$ from P_1.

Since encounters occur in a uniform manner, P_1's pay-off is equal to

$$\Pi_\mu = (b-c)q\left[\int_{\mu_{\min}}^{\mu} \mathrm{d}\mu' \int_0^{\mu'} \mathrm{d}l + \int_{\mu}^{\mu_{\max}} \mathrm{d}\mu' \int_0^{\mu} \mathrm{d}l\right] + \mathrm{b}(1-\mathrm{q}) \int_{\mu_{\min}}^{\mu_{\max}} \mathrm{d}\mu' \int_0^{\mu'} \mathrm{dl} - \mathrm{c}(1-\mathrm{q}) \int_{\mu_{\min}}^{\mu_{\max}} \mathrm{d}\mu' \int_0^{\mu} \mathrm{dl}, \tag{4.1}$$

up to normalization. The first term indicates that, when the tag is visible with probability q, the tolerance of both P_1 (μ) and the opponents (μ') must be larger than l for altruism. The second term corresponds to the situation in which P_1 sneaks benefit b without exposing that P_1 is selfish. The third term represents the opposite situation. Since $\mu_{\min} \cong 0$, equation (4.1) is reduced to

$$\Pi_\mu = -\frac{(b-c)q}{2}\left[\mu - \frac{bq-c}{(b-c)q}\mu_{\max}\right]^2 + \frac{\mu_{\max}^2}{2}\left[\frac{(bq-c)^2}{(b-c)q} + b(1-q)\right]. \tag{4.2}$$

If $bq \leq c$, the optimal tolerance is $\mu=0 \cong \mu_{\min}$, namely, unconditional defection; the same conclusion as that for the two-tag model. If $bq>c$, the optimal tolerance is given by

$$\mu = \frac{bq-c}{(b-c)q}\mu_{\max}, \tag{4.3}$$

which is smaller than $\mu_{\max}$ unless $q=1$. Now, it is beneficial to use tags. As a function of q, the best response to the uniform tolerance density is plotted in figure 4 for the three values of b/c. The optimal tolerance increases with q. Tag users prosper for a wider range of q in more benign environments (for b/c large).

We numerically simulate the continuous-tag model to confirm the above prediction. Initially, the tag and the tolerance of n players are uniformly and independently distributed on [0,1] and $[\mu_{\min}, \mu_{\max}]$, respectively. We set $n=800$, $b=1.0$ and $c=0.3$. Noiseless simulations are performed as described in §3*b*. In noisy simulations, after each generation with natural selection, eight randomly chosen players are mutated so that their tolerances are drifted by a random amount chosen uniformly from $[-0.03,0.03]$.

The population average of the tolerance after transient is shown in figure 5*a* with mutation present (filled squares) and absent (open circles). The estimation based on equation (4.3) (solid line, figure 5*a*) agrees with the stationary tolerance pretty well. Tag users can evolve for intermediate values of q, whereas extreme values of q foster unconditional defectors or generous players. The pay-off averaged over the population increases linearly with the final μ (results not shown). The standard deviations of the tolerance based on n players stay small in each run (figure 5*b*). This indicates that μ_i tends to align even under continuously distributed tags and dynamical noise.

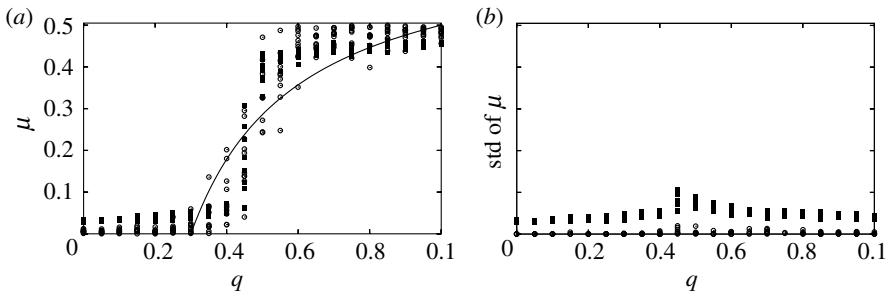

Figure 5. Results for a population of randomly interacting $n=800$ players with continuous tags and tolerances. For each q, we perform 10 runs starting with the uniform tag and tolerance densities. The duration of each run is 30 000 generations. (*a*) Population average of μ_i based on simulations with no mutation (open circles), simulations with mutation (filled squares), and the prediction by equation (4.3) (solid line). (*b*) Standard deviation of μ_i after each run.

5. DISCUSSION

We have shown that partial visibility of the tolerance (strategy) leads to tag-based indirect reciprocity. Our results extend Roberts & Sherratt's (2002) work, which showed that unconditional defection is the evolutionary outcome when the strategy of each player is completely sealed from others' eyes. If the tolerance is always observed by others, intolerant players reveal themselves to be uncooperative and cannot receive altruism from others. Thus, being generous is optimal; no room to cheat. When $bq>c$, a population in which all the players share an identical tolerance (whatever the value) is an ESS. This implies that a homogeneous population entirely of tag users with a shared tolerance level, as well as one of generous players and one of defectors, can be an ESS. Which ESS is more likely to be reached is sensitive to q, the probability that a player observes tolerance of an opponent. For an intermediate q, the unanimity of tag users results from a wide range of initial tolerance distributions including the uniform distribution. Naturally, the ultimate tolerance of the population monotonically increases with q. More generous players prosper for larger q (Nowak & Sigmund 1998*a*,*b*; Nowak *et al.* 2000; Brandt & Sigmund 2004, 2005).

The range of q hosting evolution of tag users, which is determined by b and c, is not necessarily large. However, starting from a population with more tag users than expected from the uniform density, the unanimity of tag users is more easily reached. In addition, our social lives are full of partial information. We usually have some but not complete information about others. Therefore, it is plausible to assume an intermediate value of q, for which we showed the possibility of tag-based altruism.

The mechanism that we have identified is that for indirect reciprocity, not direct reciprocity. Suppose that the player P_1 is tolerant enough to behave altruistically to player P_2. Even so, P_1 is discouraged to be nice to P_2 if P_2 is revealed to be intolerant with probability q. In other words, P_1 is not motivated to cooperate with those who will not cooperate back to P_1. Although it may sound like direct reciprocity, every decision is based on tags and tolerances, and P_1 and P_2 do not have to interact in repetition.

In the previous models of tag-based indirect reciprocity, unconditional defectors are eliminated by definition (Riolo *et al.* 2001; Traulsen & Schuster 2003). Tag users can easily evolve under this condition. By contrast, our model allows unconditional defectors. Furthermore, the boundary condition was such that random mutation somehow tended to make players more cooperative. Our model is without such a drift. A positive tolerance drift would induce a cyclic-dominance relation between different phenotypes (Riolo *et al.* 2001; Traulsen & Schuster 2003), which is not the case in our model. With unconditional defectors and without the positive drift, survival of tag users is more challenging. We have related tag-based actions to altruism by introducing the probability that tolerances are observed by others.

The proposed mechanism of indirect reciprocity does not necessitate the memory of previous acts or strategies, in contrast to image scoring (Nowak & Sigmund 1998*a*,*b*; Brandt & Sigmund 2004, 2005; Ohtsuki & Iwasa 2004) and direct reciprocity (Trivers 1971; Axelrod & Hamilton 1981; Axelrod 1984). In our model, the action depends only on present tags and tolerances. In addition, assortative mixing has small effects on the conclusions (compare figure 1*d*,*e*). Assortativity is not mandatory for our mechanism. Neither do we need population viscosity (Nowak & May 1992; Hammond & Axelrod 2006; Jansen & van Baalen 2006), which is one way to implement assortativity. Furthermore, tag-based reciprocity does not necessitate weak coupling of tag and tolerance (Jansen & van Baalen 2006). This study is a first important step towards understanding the evolution of the tag-based indirect reciprocity.

We thank Mayuko Nakamaru for helpful comments on this work. N.M. acknowledges support from the Special Postdoctoral Researchers Program of RIKEN, and H.O. acknowledges support from the Japan Society for the Promotion of Science (JSPS).

APPENDIX A.

We examine the stability of the steady-state $(p_{-1},p_0,p_1)=(1-A,0,A)$ of the symmetric two-tag model. By eliminating $(p_{-1}=1-p_0-p_1)$, we derive the two-dimensional linear dynamics around the steady state up to the second order,

$$\Delta\dot{p}_0 = \Delta p_0\left[\left(\frac{(b-c)q(1+h)}{2}+cq-c\right)\Delta p_1 + \frac{(bq-c)(1+h)}{2}\Delta p_0 + o(\Delta p_1,\Delta p_2)\right], \quad \text{(A 1)}$$

$$\Delta\dot{p}_1 = (bq-c)A\Delta p_1 + \frac{(bq-c)(1+h)A}{2}\Delta p_0 + o(\Delta p_1,\Delta p_2). \quad \text{(A 2)}$$

Here, Δp_1 and Δp_2 are perturbations to the steady state, and $o(\Delta p_1, \Delta p_2)$ summarizes small quantities relative to

Δp_1 and Δp_0. The first-order terms in the r.h.s. of equations (A 1) and (A 2) indicate that the Jacobian has eigenvalues $\lambda_1=(bq-c)A$ and $\lambda_2=0$ with corresponding eigenvectors $(p_1,p_0)=(1,0)$ and $((1+h)/2,-1)$. Provided $A>0$ (equivalently, $bq>c$), $\lambda_1>0$ follows, and the steady-state $(1-A,0,A)$ is unstable along the boundary of the simplex ($p_0=0$). When the perturbation along the first eigenmode is small ($\Delta p_1\cong 0$), the eigenmode with $\lambda_2=0$ matters for stability. When $\Delta p_1\cong 0$, equation (A 1) indicates that $\Delta\dot{p}_0>0$ inside the simplex ($\Delta p_0>0$). Therefore, the steady state is a degenerated source.

REFERENCES

Axelrod, R. 1984 *Evolution of cooperation*. New York, NY: Basic Books.

Axelrod, R. & Hamilton, W. D. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)

Bergstrom, T. C. 2003 The algebra of assortative encounters and the evolution of cooperation. *Int. Game Theory Rev.* **5**, 211–228. (doi:10.1142/S0219198903001021)

Brandt, H. & Sigmund, K. 2004 The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231**, 475–486. (doi:10.1016/j.jtbi.2004.06.032)

Brandt, H. & Sigmund, K. 2005 Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102**, 2666–2670. (doi:10.1073/pnas.0407370102)

Eshel, I. & Cavalli-Sforza, L. L. 1982 Assortment of encounters and evolution of cooperativeness. *Proc. Natl Acad. Sci. USA* **79**, 1331–1335. (doi:10.1073/pnas.79.4.1331)

Fehr, E. & Fischbacher, U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)

Hamilton, W. D. 1964 The genetical evolution of social behaviour I and II. *J. Theor. Biol.* **7**, 1–16. (doi:10.1016/0022-5193(64)90038-4) see also 17–52.

Hammond, R. A. & Axelrod, R. 2006 Evolution of contingent altruism when cooperation is expensive. *Theor. Popul. Biol.* **69**, 333–338. (doi:10.1016/j.tpb.2005.12.002)

Hofbauer, J. & Sigmund, K. 1998 *Evolutionary games and population dynamics*. Cambridge, UK: Cambridge University Press.

Jansen, V. A. A. & van Baalen, M. 2006 Altruism through beard chromodynamics. *Nature* **440**, 663–666. (doi:10.1038/nature04387)

Marsden, P. V. 1988 Homogeneity in confiding relations. *Soc. Netw.* **10**, 57–76. (doi:10.1016/0378-8733(88)90010-X)

McPherson, J. M. & Smith-Lovin, L. 1987 Homophily in voluntary organizations: status distance and the composition of face-to-face groups. *Am. Sociol. Rev.* **52**, 370–379. (doi:10.2307/2095356)

McPherson, M., Smith-Lovin, L. & Cook, J. M. 2001 Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444. (doi:10.1146/annurev.soc.27.1.415)

Newman, M. E. J. 2003 Mixing patterns in networks. *Phys. Rev. E* **67**, 026126. (doi:10.1103/PhysRevE.67.026126)

Nowak, M. A. & May, R. M. 1992 Evolutionary games and spatial chaos. *Nature* **359**, 826–829. (doi:10.1038/359826a0)

Nowak, M. A. & Sigmund, K. 1998*a* Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577. (doi:10.1038/31225)

Nowak, M. A. & Sigmund, K. 1998*b* The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574. (doi:10.1006/jtbi.1998.0775)

Nowak, M. A. & Sigmund, K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1298. (doi:10.1038/nature04131)

Nowak, M. A., Page, K. M. & Sigmund, K. 2000 Fairness versus reason in the ultimatum game. *Science* **289**, 1773–1775. (doi:10.1126/science.289.5485.1773)

Ohtsuki, H. & Iwasa, Y. 2004 How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120. (doi:10.1016/j.jtbi.2004.06.005)

Riolo, R. L., Cohen, M. D. & Axelrod, R. 2001 Evolution of cooperation without reciprocity. *Nature* **414**, 441–443. (doi:10.1038/35106555)

Roberts, G. & Sherratt, T. N. 2002 Does similarity breed cooperation? *Nature* **418**, 499–500. (doi:10.1038/418499b)

Taylor, P. D. & Jonker, L. B. 1978 Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156. (doi:10.1016/0025-5564(78)90077-9)

Traulsen, A. & Schuster, H. G. 2003 Minimal model for tag-based cooperation. *Phys. Rev. E* **68**, 046129. (doi:10.1103/PhysRevE.68.046129)

Trivers, R. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)