3B2v7.51c GML4.3.1 YJTBI : 3317 Prod.Type:COM pp.1–12(col.fig.:NIL) ED:Mangalagowri PAGN: csramesh SCAN: Vijaykumar

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Theoretical Biology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Journal of Theoretical Biology

www.elsevier.com/locate/jtbi

A tale of two defectors: the importance of standing for evolution of indirect reciprocity

Karthik Panchanathan*, Robert Boyd

Department of Anthropology, University of California, Haines Hall 341 Box 951553, Los Angeles, CA 90095, USA

Received 14 August 2002; received in revised form 2 April 2003; accepted 2 April 2003

Abstract

Indirect reciprocity occurs when the cooperative behavior between two individuals is contingent on their previous behavior toward others. Previous theoretical analysis indicates that indirect reciprocity can evolve if individuals use an image-scoring strategy. In this paper, we show that, when errors are added, indirect reciprocity cannot be based on an image-scoring strategy. However, if individuals use a standing strategy, then cooperation through indirect reciprocity is evolutionarily stable. These two strategies differ with respect to the information to which they attend. While image-scoring strategies only need attend to the actions of others, standing strategies also require information about intent. We speculate that this difference may shed light on the evolvability of indirect reciprocity. Additionally, we show that systems of indirect reciprocity are highly sensitive to the availability of information. Finally, we present a model which shows that if indirect reciprocity were to evolve, selection should also favor trusting behavior in relations between strangers.

Keywords: ■; ■; ■

1. Introduction

Thirty-one years ago Trivers (1971) laid out his theory of reciprocal altruism, which has since become the standard explanation for the evolution of cooperation between unrelated individuals. After Axelrod and Hamilton (1981) formalized Trivers' argument using an evolutionary game-theoretic model of the repeated prisoner's dilemma, myriad elaborations have been published (for example Sugden, 1986; Boyd and Lorberbaum, 1987; Boyd, 1989; Nowak and Sigmund, 1993; Boerlijst et al., 1997). Although reciprocal altruism may account for most cooperative behavior between non-kin in the animal world, new models are needed to adequately capture the complexities and subtleties of human cooperation. With some exceptions (for example Dugatkin and Wilson, 1991; Pollock and Dugatkin, 1992; Enquist and Leimar, 1993), reciprocal altruism models require repeated interaction within stable dyads ignoring processes such as communication, mobility, and reputation, which are universal features of human sociality.

Alexander (1987) suggested *indirect reciprocity* as a framework for understanding large-scale human cooperation, which "involves reputation and status, and results in everyone in the group continually being assessed and reassessed." Nowak and Sigmund (1998a, b) formalized Alexander's argument using both a game-theoretic model and computer simulations to model indirect reciprocity as a donation game in which individuals never interact with the same partner twice. In their game-theoretic model, Nowak and Sigmund (1998a) introduce reputation as an *image score* which is a state variable: individuals are either *good* or *bad*, depending on whether or not they donated in the previous round. They introduce the *Discriminator* strategy (*DISC*) which donates to those who are good and refuses to donate to those who are bad. Nowak and Sigmund find that, even under conditions in which individuals never interact more than once, the *DISC* strategy can resist invasion by indiscriminate defectors (*ALLD*). However, when indiscriminate altruists (*ALLC*) are introduced, cooperation is destabilized and defection is the only evolutionarily stable strategy (*ESS*).

In this paper, we show that indirect reciprocity cannot be based on an image-scoring strategy when errors are

*Corresponding author. Tel.: +1-301-259-1783.
E-mail address: buddha@ucla.edu (K. Panchanathan).

0022-5193/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0022-5193(03)00154-1

considered. However, if reputation is modeled as *standing* (Sugden, 1986), then indirect reciprocity can be evolutionarily stable. Like an image score, an individual's standing can either be good or bad. However, the rules governing how standings are assigned differ from those for image scores. Image scores only reflect actions: an individual that donated in the previous round will have a good image score in this round, while an individual that refused to donate will have a bad image score. Standing, on the other hand, reflects intent as well as action. So, as with image scores, any individual that donated in the previous round will have good standing. Refusals to donate, however, are parsed into two types: those that are *unjustified* and those that are *justified*. A defection is unjustified when an individual refuses to offer help to a partner in good standing. A defection is justified when an individual refuses to offer help to a partner in bad standing. While an unjustified defection always brings with it bad standing, a justified defection will leave the actor's standing unchanged.

The difference between image-score and standing is crucial. In Nowak and Sigmund's model, an individual playing the *DISC* strategy will refuse to offer help to an individual playing the *ALLD* strategy. As a result, observers will assign the *DISC* a bad image score and thus will refuse to help him in the next round. Because an image score only captures behavior and not intent, an individual is punished for refusing to help defectors. In contrast when reputation is based on standing, a refusal to help a defector does not tarnish one's reputation, only refusing help to someone in good standing will. Computing standings may require greater cognitive capacity than image scores because it requires both knowledge of others' behavior as well as inferential knowledge of their intent. The increased cognitive demands of a standing strategy may shed some light onto the evolutionary precursors to the emergence of indirect reciprocity and explain the apparent limitation of this type of behavior to humans.

In this paper, we reanalyse only Nowak and Sigmund's model (1998a) of binary-valued image scores. Their other paper on the subject (1998b) models image scores taking on values from -5 to $+5$. In that model, any donation increments one's image score by one unit, while any refusal to donate decrements one's image score by one unit. When image scores take on this range of values, rather than binary values, Nowak and Sigmund find that the image-scoring strategy, although not an ESS, can persist for long periods of time. Leimar and Hammerstein (2001), using computer simulations, argue that, even when taking on a range of values, an image-scoring strategy cannot evolve. Their criticism revolves around assumptions of population structure and the role of genetic drift. As in this paper, they find that the standing strategy is evolutionarily stable.

2. Image scoring fails when errors occur

Although not evolutionarily stable, cooperative regimes based on image scoring can persist. (Nowak and Sigmund, 1998a) When the initial frequency of image scorers is sufficiently high, selection takes the population to a mixture of image scorers and indiscriminate altruists. When this equilibrium is reached, the defectors have been driven to extinction and selection is neutral with respect to image scorers and indiscriminate altruists. Eventually, the population drifts below a critical threshold of image scorers and defectors invade and quickly go to fixation. Once this has occurred, image scorers cannot reemerge unless one considers another process such as group selection. However, Nowak and Sigmund's (1998a) claim that indirect reciprocity can be based on an image-scoring strategy hinges completely on the assumption that agents never commit errors. That is, individuals that intend to cooperate always do so and those that intend to defect always do so. In this section, by introducing errors, we show that cooperation based on image scoring cannot evolve; the defect equilibrium is reached quickly and deterministically.

We reanalyse Nowak and Sigmund's model (1998a) after adding errors. Agents interact in an infinite, unstructured population. All agents begin with an image score of good. In the first round of social interaction, each agent acts as a potential donor to a randomly selected partner. If a donation is offered, the donor's image score is good in the next round and his fitness is decremented by c while the recipient's fitness is incremented by b. It is assumed that $b > c > 0$. If no donation is offered, the image score of the donor is bad in the next round and the fitness of both the donor and the recipient remains unchanged. Subsequent rounds of social interaction occur with probability w $(0 \leqslant w < 1)$. Errors are introduced with the parameter α, which denotes the probability of an intended donation failing. We ignore the reciprocal case of an agent accidentally donating when he intended not to. Additionally we only consider errors of which both donor and recipient are aware. We distinguish these *implementation* errors from *perception* errors in which the donor believes that she donated while the recipient perceives that there was no donation. Perception errors add sufficient complexity to render analytic results intractable. However, Leimar and Hammerstein (2001) have run computer simulations considering both types of errors.

Like Nowak and Sigmund (1998a), we consider three strategies: indiscriminate altruist (*ALLC*), indiscriminate defector (*ALLD*), and the image-scoring discriminator (*DISC*). The frequencies of these strategies are denoted by x_1, x_2, and x_3, respectively. *ALLC* always donates and *ALLD* never donates. Discriminators attend to their partner's image score, donating to those

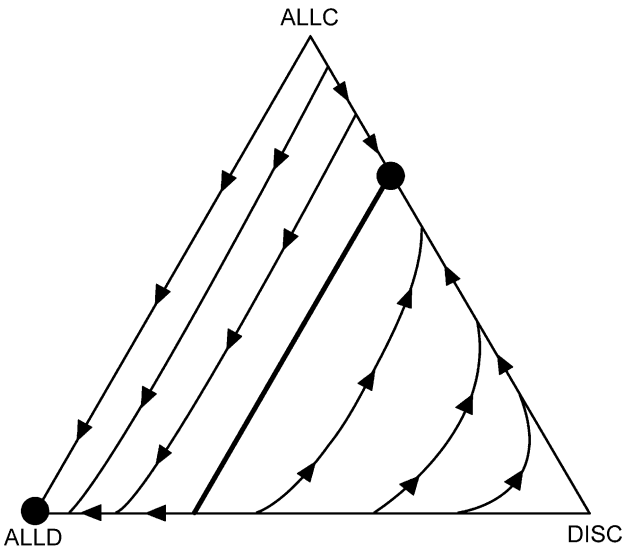

Fig. 1. Evolutionary dynamics of indirect reciprocity with errors and the *DISC* strategy. ALLD is the only ESS. Model parameters for this figure are as follows: $b = 0.01$, $c = 0.003$, $w = 0.95$, $\alpha = 0.05$. Note, the thick line represents the neutral line separating the phase space into two regions. On the right, selection leads to the *ALLC–DISC* equilibrium; on the left, selection leads to the ALLD equilibrium. Given the above parameter settings, the frequency of *DISC* (x_3) along this line is equal to 33.24% (from Eq. (4)).

who are good and withholding donations from those who are bad. The *DISC* strategy is 'nice' in that it always attempts to donate in the first round.

Employing similar analytic techniques as Nowak and Sigmund (1998a), we find that the addition of errors results in the rapid and deterministic success of defectors (Fig. 1).[1] This result is in stark contrast to the model without errors in which cooperation can persist indefinitely if the initial frequency of Discriminators is sufficiently high.

The addition of errors has a dramatic effect on the evolutionary dynamics along the *ALLC*−*DISC* edge. Recall, when errors are not considered, selection is neutral with respect to altruists and discriminators in the absence of defectors. When errors are added, there exists a stable, polymorphic equilibrium at which both *ALLC* and *DISC* are present. The frequency of *DISC* at this equilibrium is given by

$$x_3 = \frac{c}{bw(1-\alpha)}. \quad (1)$$

When errors do not occur, in the absence of defectors, everyone has a good image score and no one ever withholds donation. Thus, selection cannot distinguish between indiscriminate altruists and discriminators. When errors occur, indiscriminate altruists and discriminators respond differentially to error-committers, which then activates selection, leading to the stable, polymorphic equilibrium.

To understand this change in dynamics, first consider the case in which *ALLC* is common. The *DISC* strategy is favored by selection because it can, without cost, withhold aid from previous error-committers. When in the role of donor with a recipient who committed an error in the previous round, *ALLC* will donate, and thus have a good image score, while *DISC* will not, and thus have a bad image score. In this one interaction *ALLC* pays a cost $-c$, which *DISC* does not. In the subsequent round, as *DISC* is rare, both the good *ALLC* and the bad *DISC* will be in the role of recipient with an *ALLC* donor. As *ALLC* does not attend to image score, it will donate to either individual. The invasion criterion for *DISC* is given by

$$cw\alpha(1-\alpha) > 0. \quad (2)$$

So, if errors occur with any probability and interaction persists, *DISC* will invade.

Next, consider case in which *DISC* is common. Here, *ALLC* invades because it does not engage in costly punishment of defections from errors. When an *ALLC* finds itself in the role of donor with a recipient who had just committed an error, it donates and thus maintains its positive image score. When a *DISC* is in the role of donor to the same error-committing recipient, it does not donate and thus acquires a negative image score. Because *DISC* is common, in the subsequent round, both the good *ALLC* and the bad *DISC* will be in the role of recipient with a *DISC* donor. As such, the donor *DISC* will offer aid to the good *ALLC* and not to the bad *DISC*. In this model, discriminators are punished (i.e. they acquire a bad image score) whenever they perform their police work by withholding aid from individuals with bad image scores. *ALLC* is not punished in this way because it does not do any of the police work. The invasion criterion for *ALLC* against *DISC* is given by

$$w > \frac{c}{b(1-\alpha)}. \quad (3)$$

Finally, we look at the invasion criteria for *ALLD* along the *ALLC-DISC* edge. When condition (3) holds, there exists an unstable equilibrium given by

$$x_3 = \frac{c}{bw(1-\alpha)}. \quad (4)$$

Notice that Eq. (4) and Eq. (1) are identical. Thus, the dynamics of this system are as follows. If the initial frequency of *DISC* is below Eq. (4), *ALLD* goes to fixation. If the initial frequency of *DISC* is above Eq. (4), selection leads to the *ALLC–DISC* polymorphism given by Eq. (1). From here, any perturbation, such as drift, which momentarily decreases the frequency of *DISC*, allows *ALLD* to invade and then go to fixation.

[1] The derivation of fitness functions for each strategy and the subsequent evolutionary dynamics analyses are presented in the appendix.

Whereas cooperation through image scoring could persist indefinitely when errors are not considered, cooperation can no longer evolve when they are added.

In a world with errors and reputations of *good* or *bad*, it is no longer sufficient to know who donated and who did not. Individuals must be able to distinguish between those defections motivated by punishment and those defections motivated by selfishness. That is, strategies must parse defections into those which are justified and which are unjustified.

3. Indirect reciprocity based on standing strategies can be an ESS

Strategies based on image scoring are not evolutionarily stable because they treat all defections as negative, and as a result punish justified ones. It seems plausible that a strategy that attended to whether an observed defection is justified might be more successful. One way to do this is to introduce the notion of "standing." Everyone starts out in good standing. An individual falls into bad standing whenever he fails to cooperate with a good-standing partner (unjustified defection). A donor's standing is unchanged if he fails to cooperate with a bad-standing partner (justified defection). Good standing can be regained whenever an individual cooperates, irrespective of the standing of the partner. In this section, we test whether indirect reciprocity based on a standing strategy can evolve, looking at two variants: reputation discriminator (*RDISC*) and contrite tit-for-tat (*CTFT*). We find that indirect reciprocity based on either of these strategies can evolve. Further, these strategies are evolutionarily stable with large domains of attraction.

3.1. Reputation discrimination

The *RDISC* strategy cooperates with those in good standing and defects on those in bad standing. In this regard, *RDISC* is identical to Nowak and Sigmund's *DISC*. The difference in the two strategies lies in reputation assignment. The *RDISC* strategy assigns good standing to all those that cooperated in the previous round. When observing defections, the RDISC strategy makes a distinction between justified and unjustified defections. That is, an individual that defects on a good-standing partner is assigned bad standing while the reputation of an individual that defects on a bad-standing partner is unaltered.

As before, individuals live in an infinite, unstructured population. Each round, every individual acts as a potential donor to a randomly chosen partner. An intended donation can go wrong due to error with probability α. One round of social interaction always occurs. A subsequent round occurs with probability w.

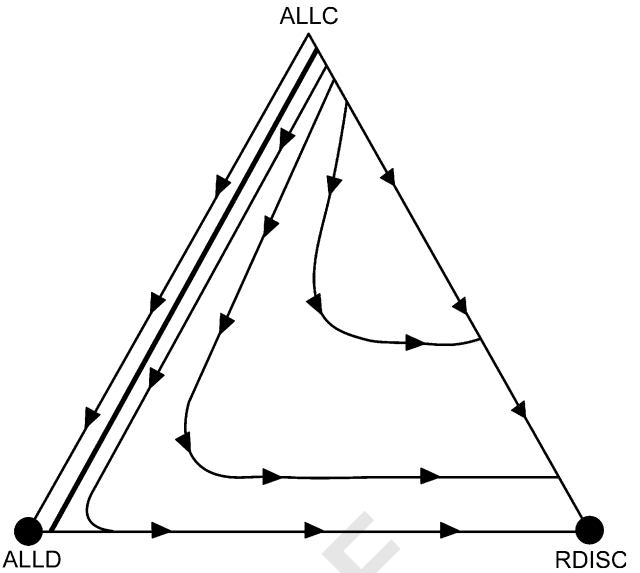

Fig. 2. Evolutionary dynamics of indirect reciprocity with the RDISC strategy. ALLD and RDISC are both stable strategies. Model parameters for this figure are as follows: $b = 0.01$, $c = 0.003$, $w = 0.95$, $\alpha = 0.05$. Note, the thick line represents the neutral line separating the phase space into two regions. On the right, selection leads to the RDISC equilibrium; on the left, selection leads to the ALLD equilibrium. Given the above parameter settings, the frequency of RDISC (x_4) along this line is approximately equal to 2.37% (from Eq. (5)).

There are three strategies: *ALLC*, *ALLD*, and *RDISC*, which have frequencies x_1, x_2, and x_4, respectively. As everyone starts out in good standing, the *RDISC* strategy always attempts to donate in the first round.[2]

As in previous models, *ALLD* is an ESS (Fig. 2). Unlike *DISC*, the *RDISC* strategy is also evolutionarily stable. Indirect reciprocity can now evolve and persist because the Reputation Discriminator strategy parses defections into those that are justified and those that are unjustified. As such, good-standing individuals that withhold donation from bad-standing individuals are not subsequently punished. Compare this with the image-scoring strategy that punishes others for withholding donation from bad-standing individuals. Also, along the entire *ALLC*−*RDISC* edge, selection favors the *RDISC* strategy. Again this is due to the ability of *RDISC* to selectively withhold cooperation from error-committers. Finally, note that the domain of attraction for *RDISC* is extensive.

Along the *ALLD*−*RDISC* edge, there is an unstable equilibrium point given by

$$x_4 = \frac{1-w}{w}\frac{c}{b-c}\frac{1}{1-\alpha}. \quad (5)$$

[2] Nowak and Sigmund (1998a) consider a *DISC* strategy that cooperates in the first round with probability p. It is shown that if p is allowed to evolve, $p = 1$ is the stable equilibrium. A similar result holds for *RDISC*.

If the initial frequency of *RDISC* is above this threshold (5), then *RDISC* increases in frequency and drives *ALLD* to extinction. If, instead, the frequency of *RDISC* is below this threshold (5), then *ALLD* increases and dominates. As in other models of cooperation, there is a minimum initial frequency of cooperators necessary to drive the defectors to extinction.

It is not possible to derive a general expression that divides the phase space between the domains of attraction for *ALLD* and *RDISC*. However, simulation results indicate that the seperatrix can be approximated using the minimum initial threshold of *RDISC* (5). If the initial frequency of *RDISC* is above (5), then the cooperative ESS usually emerges. If, instead, the initial frequency of *RDISC* is below (5), then defection dominates.

3.2. Contrite tit-for-tat

In this section, we consider the 'standing strategy' (Sugden, 1986), also known as contrite tit-for-tat (*CTFT*) (Boyd, 1989). Using computer simulations, Leimar and Hammerstein (2001) have shown that indirect reciprocity based on this strategy can evolve. *CTFT*, like *RDISC*, distinguishes between justified and unjustified defections. *CTFT* and *RDISC* use identical rules in assigning reputation. However, the two differ in their behavioral decision rules. *RDISC* only attends to its partner's standing, while *CTFT* attends to both its own standing and that of its partner. When in good standing, *CTFT* donates to those in good standing and refuses donation to those in bad standing. When in bad standing, *CTFT* always attempts to cooperate in order to regain its good standing. Otherwise, this model is identical with that presented in Section 3.1, with x_5 now denoting the frequency of *CTFT*.

The dynamics for *RDISC* (Fig. 2) and *CTFT* (Fig. 3) are similar. This is not surprising as *RDISC* and *CTFT* are similar strategies. They differ in behavior only when in bad standing, which happens to *RDISC* with probability α (9) and to *CTFT* with probability $\alpha(1-x_2)$ (14). When error rates are low probabilities Eq. (9) and Eq. (14) are close to zero. As in Section 3.1, the domain of attraction for *CTFT* is extensive.

The unstable equilibrium between *ALLD* and *CTFT* along the *ALLD*–*CTFT* edge is given by

$$x_5^3 w\alpha^2(b-c) - 2x_5^2 w\alpha(b-c) + x_5 w(b-c(1+\alpha)) - c(1-w) = 0 \quad (6)$$

which cannot be easily solved for x_5.

As in Section 3.1, a general expression for the separatrix between the domains of attraction for *ALLD* and *CTFT* cannot be derived. As before, it can be approximated using the unstable equilibrium between *ALLD* and *CTFT* (6).

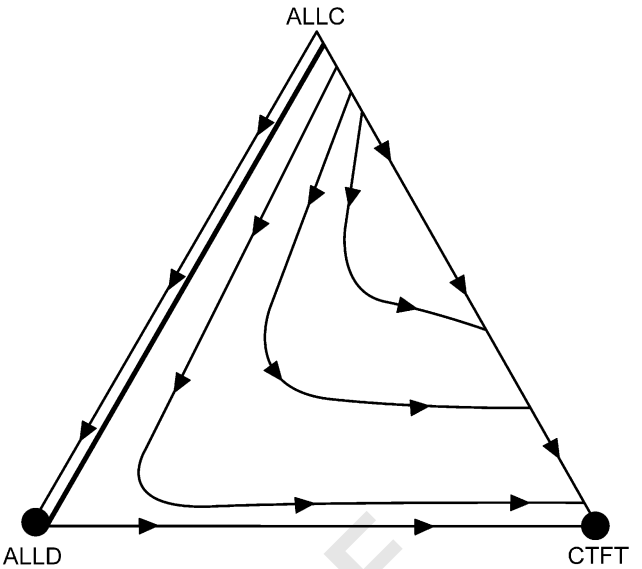

Fig. 3. Evolutionary dynamics of indirect reciprocity with the CTFT strategy. ALLD and CTFT are both stable strategies. Model parameters for this figure are as follows: $b = 0.01$, $c = 0.003$, $w = 0.95$, $\alpha = 0.05$. Note, the thick line represents the neutral line separating the phase space into two regions. On the right, selection leads to the CTFT equilibrium; on the left, selection leads to the ALLD equilibrium. Given the above parameter settings, the frequency of CTFT (x_5) along this line is approximately equal to 2.21% (from Eq. (6)).

The results from Sections 3.1 and 3.2 indicate that indirect reciprocity can be evolutionarily stable when individuals use a standing-type strategy, distinguishing between justified and unjustified defections. How individuals make this distinction and accurately infer motivation remains to be explained. Employing an image-scoring strategy, individuals need only observe the most recent action of another to determine his image score. With a standing-type strategy, individuals must also know the standings of both the donor and recipient before the action occurs. If group size is small, individuals may be able to directly observe the actions of all others. Thus, having access to the standings for both social participants, they will be able to accurately infer motivations from the most recent action. However, as group size increases, it is unreasonable to assume that individuals can directly observe the actions of all others. How does an individual interpret the actions of others for whom he has no standing information? To the extent indirect reciprocity occurs in large groups, some method of disseminating standing information seems necessary. Gossip seems to serve this function and so it appears that language must be in place before indirect reciprocity can emerge. Even with gossip, as group size increases, it is unrealistic to assume that individuals have access to the reputations of all group members. In the next section, this assumption is relaxed to test the effects of incomplete knowledge on indirect reciprocity.

4. Information availability as the limiting factor

To investigate the effect of incomplete information, we now assume that an individual knows the standing of his current partner with probability q, and with probability $1-q$ he has no information about his partner's reputation. When the partner's reputation is known *RDISC* and *CTFT* use the same decision rules outlined in the previous section. When *RDISC* and *CTFT* meet an unknown partner, we assume that they attempt to donate. In the next section we show that such trusting behavior is evolutionarily stable under a broad set of conditions. The fixed *ALLC* and *ALLD* strategies are unaffected by this new parameter.

Figs. 4 and 5 depict the separatrices between the domains of attraction for *RDISC* (Fig. 4) or *CTFT* (Fig. 5) and *ALLD* at fixed values of q. Looking at Fig. 4, suppose that q is set at 0.6. Thus, an individual knows the standing of his partner with a 60% probability. If the initial frequency of *RDISC* is sufficiently high that the population begins in the region to the right of the $q = 0.6$ line, then cooperation evolves and *RDISC* dominates. If instead, the initial frequency of *RDISC* is such that the population begins in the region to the left of the $q = 0.6$ line, then defection is the outcome; *ALLD* dominates. Note, the dividing lines in Figs. 4 and 5 are approximate; exact solutions cannot be readily derived. The unstable equilibrium along either the *ALLD*−*RDISC* edge or the *ALLD*−*CTFT* edge were used as approximate solutions for the lines separating the phase space into the regions that lead to *ALLD* and the regions that lead to *RDISC* or *CTFT*. Simulation results indicate that these approximations are very close to the exact solutions.

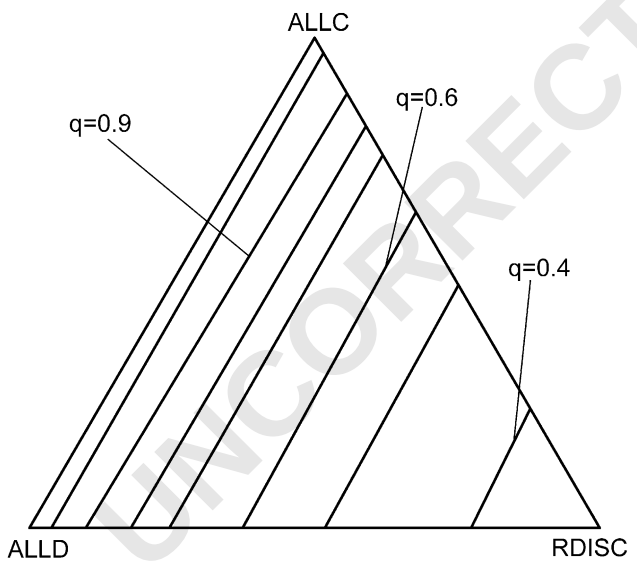

Fig. 4. Domains of attraction for ALLD and RDISC at fixed values of q. The farthest region to the left depicts the domain of attraction for ALLD at $q = 1$ (complete information). The remainder of the triangle represents the domain of attraction for RDISC at $q = 1$. Each dividing line to the right represents a decrement of 0.1 for q. The farthest region to the right depicts the domain of attraction for RDISC when $q = 0.4$. The parameter values for this figure are as follows: $\alpha = 0.05$, $w = 0.95$, $b = 0.01$, $c = 0.003$. Note, as explained in the text, these dividing lines are approximations.

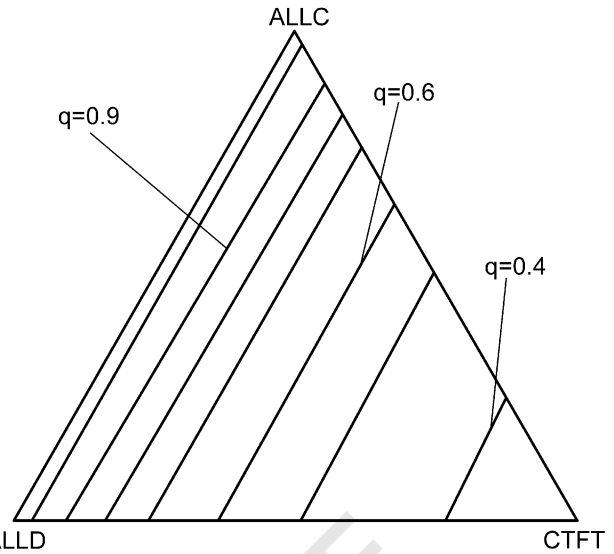

Fig. 5. Domains of attraction for ALLD and CTFT at fixed values of q. The parameter values for this figure are as follows: $\alpha = 0.05$, $w = 0.95$, $b = 0.01$, $c = 0.003$. Note, as explained in the text, these dividing lines are approximations.

From these analyses, it is clear that indirect reciprocity is highly sensitive to the degree of knowledge individuals possess with regard to others' standings. As the fraction of group members known to any individual, measured by q, decreases, the domain of attraction for the cooperative ESS rapidly diminishes.

5. Dealing with strangers and the emergence of trust

In every round of social interaction, all group members have a standing in their community. However, any particular individual may not know the standing for all other group members. In the previous section, we assumed that the *RDISC* and *CTFT* strategies always try to help strangers. We test this assumption by allowing these 'trusting' strategies to compete against 'suspicious' variants. Suspiciousness is introduced with a new model parameter, δ. Strategies intend to cooperate with partners of unknown standing with probability $1-\delta$. Thus, when $\delta = 0$, trust is complete, individuals always attempt to help strangers. We label strategies for which $\delta > 0$ as either suspicious reputation discriminator (*sRDISC*) or suspicious contrite tit-for-tat (*sCTFT*).

Figs. 6 and 7 show the conditions under which trust ($\delta = 0$) dominates suspicion in dealings with strangers. To derive these results, we first assume that cooperation

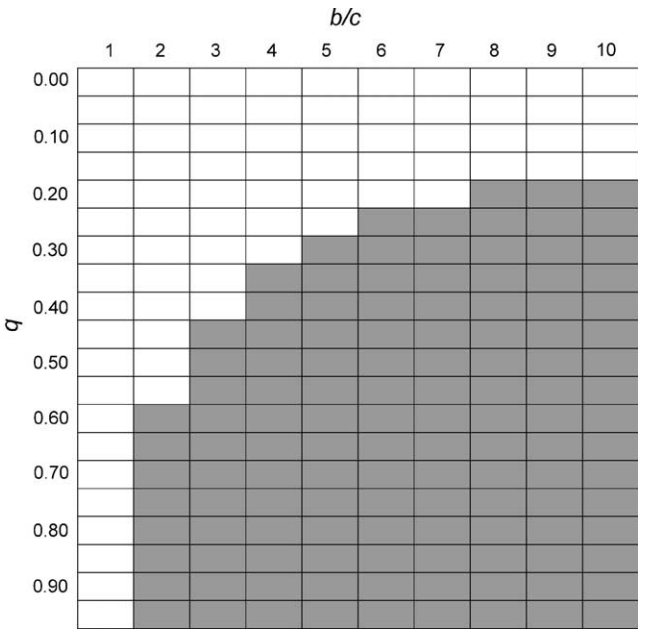

Fig. 6. Comparison of RDISC against sRDISC.

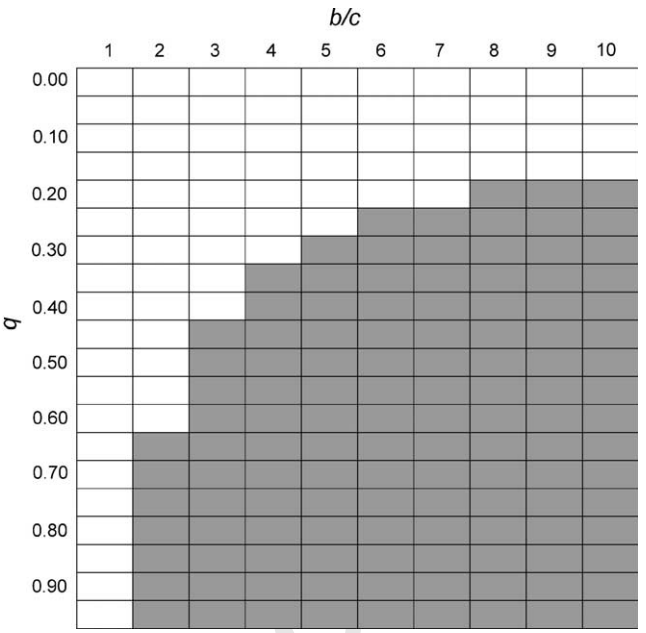

Fig. 7. Comparison of CTFT against sCTFT. In these comparisons, other strategies (*ALLC* and ALLD) are ignored. For given parameter values, we test whether the suspicious variant can invade the nice one. Note, in the first round both the nice and suspicious variants will intend to cooperate as reputations have not yet been established. After the first round, the frequencies of good-standing nice and good-standing suspicious variants rapidly converge to their respective equilibriums. Thus, for this analysis, we compare the round n payoff for the nice and suspicious variant after equilibrium has been reached. White cells represent conditions under which the suspicious variant will invade a population of the nice strategy. Grey cells represent conditions under which the nice strategy resists invasion by the suspicious variant. For these analyses, the error rate, α, was set to 0.05 and the level of suspiciousness, δ, was set to 0.01. The vertical axis represents the amount of information on reputation known to everyone, measured by the model parameter q. The horizontal axis measures the ratio of benefits of receiving cooperation to the costs of cooperating, b/c.

has evolved; the world is dominated by either *RDISC* or *CTFT*. Then, we allow suspicion to evolve by introducing *sRDISC* or *sCTFT*, setting $\delta = 0.01$. Under most conditions, trusting strategies resist invasion by suspicious ones. When the effectiveness of communication (q) is low and/or the cost of donating (c relative to b) is high, it pays to be suspicious when interacting with strangers. However, under such conditions, indirect reciprocity itself rarely emerges; defection is the more likely outcome.

6. Invasion criteria for standing strategies

Previous results indicate that indirect reciprocity based on a standing-type strategy can be evolutionarily stable. However, defection is also an *ESS*. In order for indirect reciprocity to emerge from a world of defection, cooperative individuals must assort with one another in a non-random fashion. Axelrod and Hamilton (1981) show that kin selection and reciprocal altruism operate synergistically such that with a little relatedness and a low probability of future interaction, *TFT* can invade a population of *ALLD*. Here, we determine the degree of non-random assortment (kin selection) necessary to allow for the evolution of indirect reciprocity. Under invasion conditions, cooperative individuals (*RDISC* or *CTFT*) are rare while *ALLD* is common. Thus, on average, an *ALLD* individual never meets a cooperative individual. Instead he interacts solely with other *ALLDs*. The story is different for cooperative individuals. Assuming that organisms interact preferentially with kin, there is a probability that a cooperative individual interacts with another cooperative individual because of common ancestry of the cooperative gene. This probability is measured by the model parameter r. With probability $1 - r$, the partner of a cooperative individual does not have the cooperative gene and instead plays the *ALLD* strategy.

Figs. 8 and 9 depict the relatedness required to allow for the invasion of cooperative strategies for fixed values of q and w. In order for cooperative strategies to increase when rare, there must be a high probability that individuals know the standing of one another (high values of q). The number of interactions does not appear to have a strong effect. This analysis indicates that indirect reciprocity can evolve only if social knowledge is nearly complete ($q \approx 1$). When information is complete ($q = 1$), the degrees of relatedness as a function of interaction lengths (w) necessary for cooperative strategies to increase when rare through indirect reciprocity are very close to the values found for *TFT* to increase

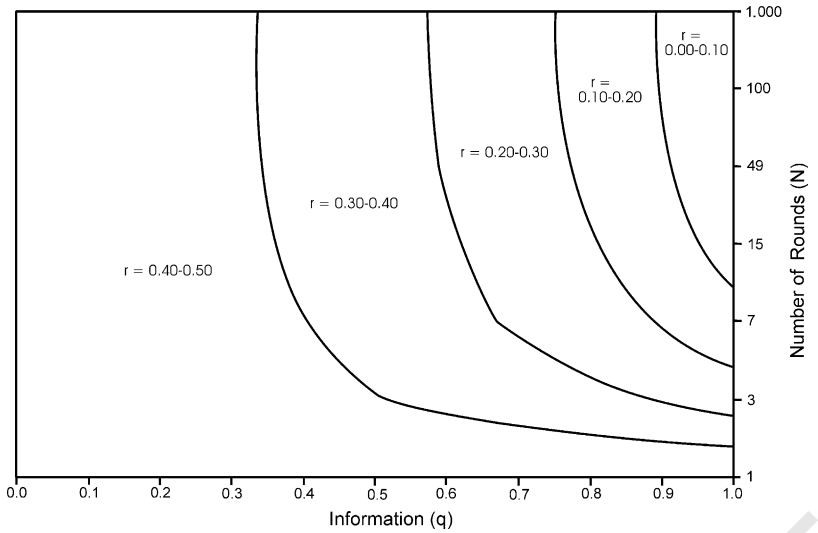

Fig. 8. Thresholds of r necessary to allow the RDISC strategy to invade ALLD. For this graph, the error rate, α, was set to 0.05. Additionally, b was set to 2 and c was set to 1. Note, Number of Rounds (N) refers to the expected number of rounds for a fixed value of w, which is given by $N = 1/1 - w$.

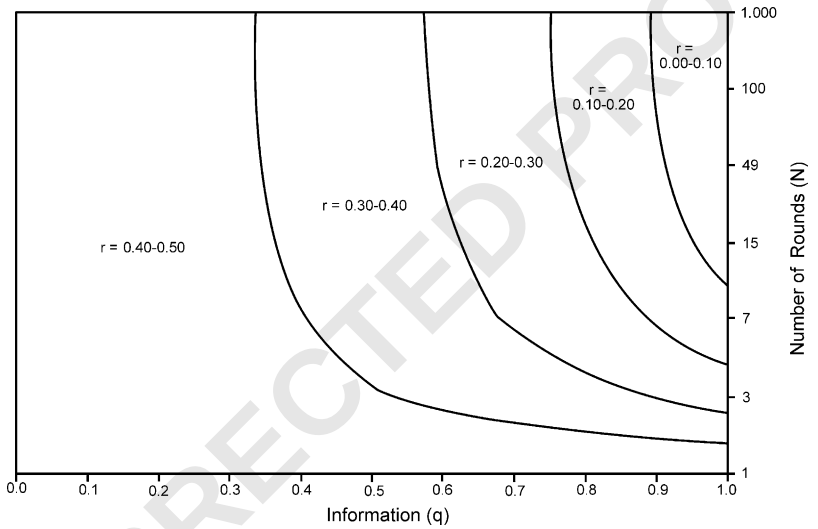

Fig. 9. Thresholds of r necessary to allow the CTFT strategy to invade ALLD. For this graph, the error rate, α, was set to 0.05. Additionally, b was set to 2 and c was set to 1. Note, Number of Rounds (N) refers to the expected number of rounds for a fixed value of w, which is given by $N = 1/1 - w$.

when rare through reciprocal altruism (as shown by Axelrod and Hamilton, 1981). However, as social knowledge tapers off (when $q \ll 1$), the conditions for the evolution of direct reciprocity and indirect reciprocity rapidly diverge.

7. *RDISC* vs. *CTFT*

In criticizing the image-scoring strategy of Nowak and Sigmund (1998b), Leimar and Hammerstein (2001) state that "image scoring strategies fail to represent the true strategic interests of an individual. Individuals cannot benefit by decisions based partly or wholly on the score of a potential recipient. The only influence of an individual's current aid-giving decision on the probability of receiving aid in the future is due to the change in the individual's own score. A rational individual in this setting should then use a strategy that takes his or her own score into account, but ignores the score of a potential recipient". In this section, we test this assertion by comparing the fitness of the two standing-type strategies thus far analysed (*RDISC* and *CTFT*) when they interact with each other. By setting the frequencies of *ALLD* and *ALLC* to zero, we can now look at which standing strategy variant selection favors.

Under all parameter settings, *CTFT* has marginally higher fitness than *RDISC*. This slight fitness differential is attributed to the *CTFT* decision rule of cooperating when in bad standing. Recall, *RDISC* only cooperates when matched with a good-standing partner. Because this fitness differential is small, selection between the two strategies is essentially neutral. In such cases, other processes determine which strategy ultimately flourishes. First, perception errors have not been considered. The two strategies may respond differentially to these types of errors. Second, we have thus far stated that both *RDISC* and *CTFT* look favorably upon all cooperative acts. If the *RDISC* strategy were altered such that cooperation with bad standing individuals confers bad standing on the donor, *CTFT* would not be able to invade. This decision rule may approximate cultural norms that frown upon upstanding members of the community associating with known felons.

8. Discussion

To understand the evolution of human cooperation, new models are needed that move beyond reciprocal altruism. The model of indirect reciprocity presented by Nowak and Sigmund (1998a) represents an important step in this direction. Although dealing with an extreme and simplified case: a donation game with new partners every round, Nowak and Sigmund's model touches upon unique aspects of human sociality that previous models were not able to capture such as trust, gossip, and reputation (for exceptions see Dugatkin and Wilson, 1991; Pollock and Dugatkin, 1992; Enquist and Leimar, 1993). However, as we demonstrate in this paper, if a process such as indirect reciprocity were to evolve and reputations were binary (either be *good* or *bad*), it could not be based on an image-scoring strategy (see also Leimar and Hammerstein, 2001). Image scoring requires only that agents be able to acquire information as to the actions of others. Those that cooperate are assigned positive image scores, while those that defect are assigned negative image scores. Our analysis shows that additional information is necessary to stabilize indirect reciprocity. Specifically, individuals must be able to infer motivations from observed defections, parsing them into those that are justified and those that are unjustified. Strategies that use standing have this property and are found to be evolutionarily stable.

In this paper, we analyse two standing strategy variants (*RDISC* and *CTFCT*) to show that it is not crucial that individuals attend to their own standing when deciding upon a course of action. However, it is crucial that individuals are able to discern motivation from observed defection. The question then becomes how individuals make this distinction. In systems of reciprocal altruism, organisms can track the actions of their partners through direct observation. When one partner defects out of turn, the other can take appropriate action. In a system of indirect reciprocity, where individuals are constantly paired with new exchange partners, what does an individual make of a partner who defected on his previous partner? Was the defection retaliatory or selfish? If group size is small, perhaps individuals can monitor the goings on of all others and thus properly attribute standing to a partner observed defecting on another. As group size increases, however, this assumption seems implausible. Language seems to offer individuals access to information about others that they were not able to observe directly. Integrating this hearsay with personally observed information, individuals may be able to accurately track the standings of other group members. This argument is parsimonious with the observation that something like indirect reciprocity seems to be extremely rare in nature with the notable exception of humans. Without effective communication, reputations may only exist in the context of a stable dyad. Once communication is possible, an individual's reputation takes on more global characteristics.

Recent experiments investigate the propensity of subjects to cooperate when there is no expectation of reciprocation from current partners. Wedekind and Milinski (2000) find that individuals playing a donation game are more likely to donate to those that had previously donated. However, as Leimar and Hammerstein (2001) have noted, these experiments do not allow for a standing strategy to be implemented. It is unclear whether potential donors would respond differentially to observed justified and unjustified defections. To test this, Milinski et al. (2001) add second-order information with the hypothesis that image-scoring strategists would not use this second-order information whereas standing strategists would. The authors find that subjects are as likely to withhold donation from justified defectors as they are from unjustified defectors, which is indicative of an image-scoring strategy. This interpretation is misleading because it fails to capture how the standing strategy works in practice. The model presented by Milinski et al. (2001) assumes that the standing strategy uses entire action-histories for all other group members to infer motivation from recently observed actions. The authors ignore language, the crucial adaptation that makes indirect reciprocity possible. Rather than encoding action-histories, individuals encode standings (in this case a state variable with a value *good* or *bad*) for other group members, engage in social interaction, communicate with one another, and then update standings. An experiment that incorporates this would allow one to accurately distinguish between an image-scoring strategy and a standing strategy.

It is interesting to note that Nowak and Sigmund (1998a) mention the standing strategy as possibly a better candidate for explaining indirect reciprocity. They did not model the strategy stating that it would be susceptible to errors in perception. Including such errors adds sufficiently complexity to render analytic techniques intractable. However, Leimar and Hammerstein (2001) subjected the standing strategy to errors in perception in their computer simulations. So long as errors in perception are sufficiently low relative to errors in implementation, the standing strategy is an *ESS*. Preliminary work by one of the present authors (Panchanathan) suggests that when agents are embedded in social networks, this restriction is lifted; cooperation based on a standing strategy can be sustained in the face of high rates of perception error.

The differential response of the image-scoring and the standing strategy to implementation errors, in which both partner and recipient are aware of an error occurring, is worth mentioning. In the face of implementation errors, at any rate, the image-scoring strategy has no success. The population rapidly converges to an all defect equilibrium. (Fig. 1) The standing strategies (*RDISC* and *CTFT*) are evolutionarily stable in the face of relatively high rates of implementation errors. (Figs. 2 and 3) In fact, errors stabilize cooperation. Analogous results are found in Boyd (1989) in a model of reciprocal altruism. In the absence of defectors, selection cannot distinguish between a standing strategy and the unconditionally cooperative strategy unless errors occur at some rate. Lotem et al. (1999) discuss this extensively in their model of "phenotypic defectors".

Even if individuals live in small communities, there are going to be instances when strangers are forced to interact. In such interactions, how should individuals behave? When cooperation has been established, our results indicate that a 'kindness to strangers' norm is optimal (for a model in which 'suspiciousness' evolves, see Enquist and Leimar (1993)). As reputations spread quickly, it pays to help a stranger, even if it turns out that he is a career defector. By so doing, one hopes to raise one's own standing in the community and thus be a target of future generosity.

Acknowledgements

We would like to thank Dan Fessler, Alan Fiske, Siamak Naficy, Karl Sigmund, the UCLA Biological Anthropology Modeling Group and three anonymous reviewers for their thoughtful comments. Karthik Panchanathan was funded by an NSF Graduate Research Fellowship. This paper was first presented at the Human Behavior and Evolution Society (HBES) conference in June, 2001.

Appendix A. Image scoring fails when errors occur

In this section we generalize the model of indirect reciprocity presented by Nowak and Sigmund (1998a) by including errors. The error-rate is denoted by the parameter α. Two simplifying assumptions are made. First, only one-way errors can occur (intended cooperation leading to accidental defection). Second, errors are assumed to be sufficiently rare that the possibility of simultaneous errors can be ignored.

The frequencies of the strategies *ALLC*, *ALLD*, and *DISC* are denoted by x_1, x_2, and x_3, respectively. The parameter w measures the probability of an additional round of social interaction. The parameter b measures the incremental benefit to fitness received by an individual when his partner cooperates. Likewise, c measures the incremental cost borne by the actor when he cooperates.

Employing similar methods to Nowak and Sigmund (1998a), the following fitness functions can be derived:

$$W(ALLC) = \frac{1-\alpha}{1-w}[bx_1 + bx_3(1-w\alpha) - c],$$
$$W(ALLD) = \frac{1-\alpha}{1-w}bx_1 + (1-\alpha)bx_3, \quad (7)$$
$$W(DISC) = \frac{[b - c(1-\alpha) - bx_3w\alpha(1-\alpha)][1-w+wx_1(1-\alpha)]}{(1-w)(1-wx_3(1-\alpha))} - b[x_2 + \alpha(x_1+x_3)].$$

From these fitness functions, the derivations of Eqs. (1)–(4) are straightforward. To find the equilibrium between *ALLC* and *DISC* (1), simply find the value of x_3 when $W(ALLC) = W(DISC)$, setting $x_1 = 1 - x_3$ and $x_2 = 0$. For the invasion criterion of *DISC* against *ALLC* (2), set $x_3 = x_2 = 0$ and $x_1 = 1$, finding the condition under which $W(DISC) > W(ALLC)$. For the invasion criterion of *ALLC* against *DISC* (3), set $x_1 = x_2 = 0$ and $x_3 = 1$, finding the condition under which $W(ALLC) > W(DISC)$. Finally, for the invasion criterion of ALLD against a population comprised of *ALLC* and *DISC* (4), we find when $W(ALLD) > x_1W(ALLC) + x_3W(DISC)$ given $x_1 = 1 - x_3$ and $x_2 = 0$. When condition (3) holds, the unstable equilibrium (4) is the only one in the interval of [0,1]. When (3) does not hold, the shadow of the future is not sufficiently large; ALLD is the only ESS.

A.1. Indirect reciprocity based on standing strategies can be an ESS

Here, we outline the model of indirect reciprocity with the *RDISC* and *CTFT* strategies. We only present the results of the most general model, which includes incomplete information. This is introduced by the

parameter q, which measures the probability that an individual knows the standing of his current partner.

A.2. Reputation discrimination

The frequencies for *ALLC*, *ALLD*, and *RDISC* are given by x_1, x_2, and x_4, respectively. It is also necessary to track the frequency of individuals for a given strategy in good standing in any particular round. We denote these frequencies with $g_n(ALLC)$, $g_n(ALLD)$, $g_n(RDISC)$, which refer to the frequency of good standing individuals in round n for *ALLC*, *ALLD*, and *RDISC*, respectively. Additionally, we denote the frequency of good-standing individuals among the whole population as g_n. In round 1, as no interactions have yet taken place, we assume all individuals are in good standing.

Thus in round n, where $n > 1$,

$$\begin{aligned} g_n(ALLC) &= g_{n-1}(ALLC)(1 - g_{n-1}\alpha) \\ &\quad + (1 - g_{n-1}(ALLC)) \\ &\quad \times (1-\alpha) g_n(ALLD) = 0, \end{aligned}$$

$$\begin{aligned} g_n(RDISC) &= g_{n-1}(RDISC)(1 - g_{n-1}\alpha) \\ &\quad + (1 - g_{n-1}(RDISC)) \\ &\quad \times k(1-\alpha)[q g_{n-1} + (1-q)], \end{aligned}$$

$$g_n = x_1 g_n(ALLC) + x_4 g_n(RDISC). \tag{8}$$

As these recursions cannot be solved, we cannot compute fitness functions directly. However, simulation results indicate that these recursions rapidly converge. Therefore, to derive the fitness functions, we add the first round payoffs to the payoffs in round n, by which the recursions in Eq. (8) have converged, weighted by $w/1-w$. This last term reflects the weighting given to all rounds after the first. Simulation analysis reveals that little is lost using this approach.

Before computing fitnesses, we derive the equilibria for the recursions in Eq. (8).

$$\begin{aligned} &g_n(ALLC) \to 1 - \alpha(1 - x_2), \\ &g_n(RDISC) \to 1 - \frac{\alpha(1-x_2)}{1 - q x_2}. \end{aligned} \tag{9}$$

To find these equilibria, we make the following approximations. At equilibrium, these frequencies are close to 1. Thus all terms such as $(1 - g_n(ALLC))^2, \alpha(1 - g_n(ALLC))$, and α^2 tend to zero. The same holds true for *RDISC*.

To compute fitness functions, we derive both round 1 payoffs and the payoffs in round n, where condition (8) has been reached.

$$\begin{aligned} &W_1(ALLC) = (1-\alpha)[b x_1 + b x_4 - c], \\ &W_1(ALLD) = (1-\alpha)[b x_1 + b x_4], \\ &W_1(RDISC) = (1-\alpha)[b x_1 + b x_4 - c], \end{aligned} \tag{10}$$

$$\begin{aligned} W_n(\mathrm{ALLC}) &= (1-\alpha) \\ &\quad \times [b x_1 + b x_4 (q g_n(\mathrm{ALLC}) + (1-q)) - c], \\ W_n(\mathrm{ALLD}) &= (1-\alpha)[b x_1 + b x_4 (1-q)], \\ W_n(\mathrm{RDISC}) &= (1-\alpha)[b x_1 + b x_4 (q g_n(\mathrm{RDISC}) \\ &\quad \times + (1-q)) - c(q g_n + (1-q))]. \end{aligned} \tag{11}$$

Summing up:

$$\begin{aligned} &W(ALLC) \\ &\quad = \frac{1-\alpha}{1-w}[b x_1 + b x_4 (1 - wq(1 - g_n(\mathrm{ALLC})) - c], \\ &W(\mathrm{ALLD}) = \frac{1-\alpha}{1-w}[b x_1 + b x_4 (1 - wq)], \end{aligned} \tag{12}$$

$$\begin{aligned} W(\mathrm{RDISC}) = \frac{1-\alpha}{1-w}[&x_1\{b - c(1 - wq(1 - g_n(ALLC)))\} \\ &+ x_2\{-c(1-wq)\}] + \frac{1-\alpha}{1-w} \\ &[x_4(b-c)\{1 - wq(1 - g_n(RDISC))\}]. \end{aligned}$$

To find the unstable equilibrium between *RDISC* and *ALLD* (5), set $W(ALLD) = W(RDISC)$, $x_2 = 1 - x_4$ and $q = 1$.

A.3. Contrite tit-for-tat

The frequencies for *ALLC*, *ALLD*, and *CTFT* are given by x_1, x_2, and x_5, respectively. As before, we track the frequency of individuals for a given strategy in good standing in any particular round. These frequencies are denoted by $g_n(ALLC), g_n(ALLD), g_n(CTFT)$, which refer to the frequency of good standing individuals in round n for *ALLC*, *ALLD*, and *CTFT*, respectively. Additionally, we denote the frequency of good-standing individuals among the whole population as g_n.

$$\begin{aligned} g_n(ALLC) &= g_{n-1}(ALLC)(1 - g_{n-1}\alpha) \\ &\quad + (1 - g_{n-1}(ALLC))(1-\alpha) g_n(ALLD) = 0, \end{aligned}$$

$$\begin{aligned} g_n(CTFT) &= g_{n-1}(CTFT)(1 - g_{n-1}\alpha) \\ &\quad + (1 - g_{n-1}(CTFT))(1-\alpha), \end{aligned} \tag{13}$$

$$g_n = x_1 g_n(ALLC) + x_5 g_n(CTFT).$$

Using similar approximations as before, we find the equilibria for these recursions.

$$\begin{aligned} &g_n(ALLC) \to 1 - \alpha(1 - x_2), \\ &g_n(CTFT) \to 1 - \alpha(1 - x_2). \end{aligned} \tag{14}$$

As before, we find the payoffs in round 1.

$$\begin{aligned} &W_1(ALLC) = (1-\alpha)[b(x_1 + x_5) - c], \\ &W_1(ALLD) = (1-\alpha) b(x_1 + x_5), \\ &W_1(CTFT) = (1-\alpha)[b(x_1 + x_5) - c]. \end{aligned} \tag{15}$$

In round n we have

$$W_n(ALLC) = (1-\alpha)[bx_1 + bx_5(g_n(CTFT) \times [qg_n(ALLC) + (1-q)] + (1-g_n(CTFT))) - c],$$

$$W_n(ALLD) = (1-\alpha)[bx_1 + bx_5(g_n(CTFT)(1-q) + (1-g_n(CTFT)))], \quad (16)$$

$$W_n(CTFT) = (1-\alpha)[bx_1 + bx_5(g_n(CTFT)[qg_n(CTFT) + (1-q)] + (1-g_n(CTFT)))] - (1-\alpha)[c(g_n(CTFT)[qg_n + (1-q)] + (1-g_n(CTFT)))].$$

Summing up:

$$W(ALLC) = \frac{1-\alpha}{1-w}[bx_1 + bx_5(1 - wqg_n(CTFT)(1-g_n(ALLC))) - c],$$

$$W(ALLD) = \frac{1-\alpha}{1-w}[bx_1 + bx_5(1 - wqg_n(CTFT))], \quad (17)$$

$$W(CTFT) = \frac{1-\alpha}{1-w}[bx_1 + bx_5 \times (1 - wqg_n(CTFT)(1-g_n(CTFT)1 - wqg_n(CTFT)(1-g_n)) - c].$$

To find the unstable equilibrium between *CTFT* and *ALLD* (6), set $W(ALLD) = W(CTFT)$, $x_2 = 1 - x_5$ and $q = 1$.

References

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

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.

Boerlijst, M.C., Nowak, M.A., Sigmund, K., 1997. The Logic of Contrition. J. Theor. Biol. 185, 281–293.

Boyd, R., 1989. Mistakes Allow Evolutionary Stability in the Repeated Prisoner's Dilemma Game. J. theor. Biol. 136, 47–56.

Boyd, R., Lorberbaum, J.P., 1987. No pure strategy is evolutionarily stable in the repeated prisoners' dilemma game. Nature 327 (6117), 58–59.

Dugatkin, L.A., Wilson, D.S., 1991. Rover: A strategy for exploiting cooperators in a patchy environment. Am. Nat. 138 (3), 687–701.

Enquist, M., Leimar, O., 1993. The evolution of cooperation in mobile organisms. Anim. Behav. 45, 747–757.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond. B-Biol. Sci. 268 (1468), 745–753.

Lotem, A., Fishman, M.A., Stone, L., 1999. Evolution of cooperation between individuals. Nature 400 (6741), 226–227.

Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.J., 2001. Cooperation through indirect reciprocity: Image scoring or standing strategy? Proc. Roy. Soc. Lond. Ser. B-Biol. Sci. 268 (1484), 2495–2501.

Nowak, M.A., Sigmund, K., 1993. A strategy of win stay, lose shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature 364 (6432), 56–58.

Nowak, M.A., Sigmund, K., 1998a. The dynamics of indirect reciprocity. J. Theor. Biol. 194 (4), 561–574.

Nowak, M.A., Sigmund, K., 1998b. Evolution of indirect reciprocity by image scoring. Nature 393 (6685), 573–577.

Pollock, G., Dugatkin, L.A., 1992. Reciprocity and the emergence of reputation. J. Theor. Biol. 159 (1), 25–37.

Sugden, R., 1986. The Economics of Rights, Cooperation and Welfare. Basil Blackwell, Oxford.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Q. Rev. Biol. 46, 35–57.

Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288 (5467), 850–850.