Evolutionary Pathways to Group Cooperation: Reputation, Reciprocity, and Signaling

Karthik Panchanathan (Anthropology, UCLA)

Eric Alden Smith (Anthropology, U of Washington)

[Version 8.1, 1 June 2010]

1. Introduction

The problem of cooperation is a central issue in the biological and social sciences. How can natural selection (or alternatively, cultural evolution or decision-making by self-interested individuals) produce patterns of cooperation that involve helping others at a cost to the actor? Darwin (1859, 1871) struggled with this "special difficulty," as exemplified by the reproductive division of labor among colonial insect species, and the group-beneficial morality and fellowship found in human societies, and worried that it might be "actually fatal to the whole theory" of adaptation by natural selection. A century passed before biologists offered detailed solutions to Darwin's "special difficulty." Hamilton's (1964) "inclusive fitness" concept re-imagined natural selection acting at the level of alleles, rather than the individuals housing those alleles, leading to the theory of "kin selection" (Maynard Smith 1964; see also Williams and Williams 1957). Subsequently, reciprocity theory offered an explanation for cooperation between unrelated individuals (Trivers 1971; see also Taylor 1976, Schaffer 1978, and Axelrod and Hamilton 1981).

A quarter century ago, the consensus among evolutionists was that some combination of kin selection and reciprocity sufficiently explained all instances of cooperation—from sterile insect castes to baboon alliances, from blood sharing between vampire bats to complex forms of human cooperation such as warfare and big-game hunting. The story is no longer so simple. Though kin selection and reciprocity still feature prominently as explanatory mechanisms for the evolution of cooperation, a growing number of specialists, especially those interested in human evolution, view them as insufficient. Indeed, in the cases of greatest anthropological interest, involving group cooperation (more than 2 or 3 individuals) between individuals of low average relatedness, these two classical models seem inadequate—a situation not widely appreciated by non-specialists. A variety of alternative explanations for group cooperation have been proposed, including punishment, reputation, costly signaling, and group selection (Boyd et al. 2003; Fehr and Gachter 2002; Milinski et al. 2002; Gintis et al. 2001). While a consensus has not yet emerged, a robust explanation will certainly include some role for reputation. In this paper, we explicate key models of the evolution of cooperation that include reputation, discuss their value as well as their limitations in accounting for group cooperation, and attempt to clarify some of the ambiguities and terminological confusions that have arisen in this area of research.

2. Group cooperation: Scale and assortment

The problem of collective action

All social species face social dilemmas, in which achieving cooperation is difficult because the best move from a selfish point of view does not produce the best collective outcome. In the social sciences, these are often termed "collective action problems", situations where several or many individuals must cooperate in order to produce some collective good (see Box 1 for a glossary of these and other terms). Collective goods abound in every human society, including large game that is widely shared, community irrigation systems, defense against enemy attack, public education, and clean air. Other social species produce collective goods as well, but if we restrict ourselves to cases where the members of the group are not closely related (thus ruling out kin selection as the dominant explanatory mechanism), the number of non-human examples dwindles drastically.

[INSERT BOX 1 HERE]

Successful collective action is difficult to achieve for at least two reasons. There may be logistical or informational constraints on getting the relevant individuals to act in concert—what game theorists term a coordination problem. Or—the more interesting case—there may be insufficient incentives to motivate individuals to contribute. This second situation constitutes a collective action problem, meaning any case where individuals have an incentive to "free ride" by refusing to contribute to while still partaking in the benefits of collective action (Olson 1965), or to selfishly consume more than one's share of a collective good—the infamous "tragedy of the commons" (Hardin 1968). The incentives to free ride or over-appropriate often increase with group size, as each individual's impact is spread over a larger number, and the chances of being caught decline. That humans readily solve collective action problems in large groups (Ostrom 1990), despite low relatedness, demands an explanation.

Assortment: Kinship and beyond

Kinship is one pathway to costly cooperation. Cooperation can evolve if the marginal fitness gain to the recipient devalued by the degree of relatedness is greater than the marginal fitness cost to the donor. This result is captured by Hamilton's rule: $rb - c > 0$, where r measures relatedness, and b and c refer to the recipient's marginal benefit and the donor's marginal cost, respectively. Kin selection works because an allele for altruism (to put it crudely) can spread if it has the effect of enhancing the fitness of copies of this allele found in other bodies (Grafen 1984). Seen this way, kinship is a device for positive assortment of cooperators: high relatedness increases the chance that altruism will enhance inclusive fitness, hence favoring the spread of the allele coding for it.

Given the facts of human reproduction and demography, relatedness seems too low to explain the evolution of group cooperation in humans. But kinship is just one possible assortment device. There is scope for the evolution of cooperation whenever cooperators can selectively channel cooperation to fellow cooperators, whatever the mechanism that generates the assortment. In direct reciprocity, past behavior is used as an assortment device: by matching cooperation for cooperation and defection for

defection, the benefits of cooperation selectively flow to those who cooperate. While the assumptions of direct reciprocity are few and unrestrictive, direct reciprocity seems to be rare outside of humans (Hammerstein, 2003). Further, while direct reciprocity can be a powerful mechanism in generating dyadic cooperation, it does a poor job of explaining group cooperation (Boyd and Richerson, 1988), a point to which we will return. In models of indirect reciprocity, actors condition their behavior on their partners' past dealings with third-parties. By selectively channeling cooperation to those who cooperate with third-parties, reputation becomes an assortment device that can drive the evolution of cooperation. These kinds of models are extremely similar to models of direct reciprocity: instead of conditioning your behavior on your partners' past behavior toward you, you condition your behavior on your partners' past behavior towards others. In signaling models, only individuals of high quality can engage in costly cooperation. Because audience members can reliably infer quality (a latent variable) from behavior (a conspicuous act), reputation for generosity serves as an assortment device whereby audience members selectively interact with high quality individuals. Indirect reciprocity and costly signaling represent at least two ways in which reputation has been invoked to explain the evolution of cooperation. These two mechanisms share many fundamental elements, even as they chart somewhat alternate routes to cooperation.

3. Indirect reciprocity and dyadic cooperation

In his seminal work on direct reciprocity, Trivers (1971) suggested selection might have broadened reciprocity to encompass groups larger than a dyad, an outcome he called "generalized altruism." Alexander (1987), developing this idea, argued that systems of "indirect reciprocity" underlie human morality and play a role in shaping the evolution of cooperation. Nowak and Sigmund (1998) formalized at least one part of Alexander's verbal argument, showing how indirect reciprocity might offer an explanation of some aspects of cooperation between non-kin. In this model, individuals interact with strangers and choose whether or not to cooperate based on their partners' reputation. Reputation represents a summary of the partner's past behavior toward third parties, as interpreted by the community or some aggregate set of accounts.

To see how this model works, we describe some of the key features and highlight the results (following Panchanathan and Boyd 2003). We start with a large (actually infinite) population. Individuals pair up at random; engage in a one-shot social exchange, in which each individual can provide a benefit b to his partner, at a personal cost c; and, with a fixed probability w, individuals form new pairs and engage in another social exchange. Each social exchange is a prisoner's dilemma: an individual's welfare is maximized by not helping his partner, whereas the group's welfare (here, the dyad's) is maximized when help is proffered (i.e., $b > c > 0$).

This model is similar to Axelrod and Hamilton's (1981) model of direct reciprocity, but for one crucial difference. In direct reciprocity, with a fixed (population-wide) probability w, a dyad persists and individuals interact with the same partner—Axelrod's (1984) "shadow of the future"; with a fixed probability $1–w$, the dyad breaks up and social interaction ends. In indirect reciprocity, dyads always break up after social interaction. With a fixed probability w, a new dyad is formed and individuals engage in another one-shot prisoner's dilemma; with a fixed probability $1–w$, new dyads are not formed and social interaction ends. In models of indirect reciprocity,

individuals never interact with the same partner twice. This feature of indirect reciprocity models is not meant to reflect reality; it is an idealization that is meant to isolate the effects of reputation from direct reciprocity.

So, if individuals only interact with strangers, never with previous exchange partners, what motivates cooperation? With a fixed probability q, an individual knows her partner's reputation, which is a summary representation of that individual's past dealings with third parties. Although there are many ways to summarily represent a sequence of social exchanges (e.g., Ohtsuki and Iwasa 2006), indirect reciprocity models for the most part assume that all community members share the same rule. For this paper, following Sugden (1986; Panchanathan and Boyd, 2003), we consider a reputation rule in which individuals remain in "good" standing so long as they don't defect on partners who are also "good". For mathematical tractability, certain assumptions are made about the dissemination and accuracy of reputations, including consensus, such that all community members agree on the reputation of a particular individual. Whether such assumptions are valid are, of course, empirical questions.

As in direct reciprocity models, there are two behavioral strategies considered in most models of indirect reciprocity: *reciprocators* help partners of good or unknown standing, whereas *defectors* never cooperate. *Reciprocator* is an evolutionarily stable strategy, such that a population composed of them is resistant to invasion by *defectors*, provided the following condition is satisfied:

$$qwb > c \qquad (1)$$

The term to the right of the inequality sign represents the cost of helping. *Reciprocators* pay this cost to help partners of good or unknown standing, while *defectors* never pay this cost. When *reciprocators* are common, the average *reciprocator* will only interact with other *reciprocators*, so this cost will always be paid. The term to the left of the inequality sign represents the incremental benefit *reciprocators* enjoy over *defectors*. *Reciprocators* receive this benefit on the subsequent round of social interaction by helping their partner in the current round. This incremental benefit is the product of three terms: the probability that another round of social interaction occurs (w); the probability that recipient's reputation is known (q; because *reciprocators* help partners of unknown reputation, there is no selective benefit for *reciprocators* when reputations are unknown as *defectors* will receive help from other *reciprocators*, too); and the benefit of receiving help (b).

When each community member knows the reputations of all the others (q= 1), Condition (1) reduces to the same stability condition Axelrod and Hamilton (1981) found for the *tit for tat* strategy in their model of direct reciprocity. Upon reflection, this shouldn't be surprising. Reciprocity works when individuals can strategically condition their behavior on their partners' past behavior. In direct reciprocity, *tit for tat* players channel cooperation to those who previously cooperated with them and withhold from those who did not, thereby resisting invasion by *defectors*; in indirect reciprocity, despite ephemeral relationships, the same logic applies: *reciprocators* channel cooperation to those who have been cooperative with third parties and withhold from those who have not, thereby resisting invasion by *defectors*. As individuals track a smaller and smaller subset of the goings on of others (q < 1), the prospect of cooperation through indirect reciprocity diminishes. For small groups, like hunter-gatherer bands

or friendship *cliques*, gossip might be sufficient to broadcast reputations and stabilize cooperation through indirect reciprocity; as group sizes increase to the scale of large chiefdoms or states, without some institutional mechanism, such as provided on eBay, indirect reciprocity by itself is unlikely to maintain cooperation (for an historical example of such a mechanism, see Greif 1989).

It is not enough to verify that cooperation via a purported mechanism is evolutionarily stable; a complete explanation must also demonstrate that cooperation can plausibly evolve when initially rare. After all, any evolutionary explanation of cooperation presupposes an ancestral, uncooperative state, and a general feature of reciprocity models is that these uncooperative equilibria are evolutionarily stable, posing the problem of how cooperative strategies can initially spread. Following Maynard Smith (1982), we determine whether cooperation can evolve when initially rare by assuming that mutant *reciprocators* are exceedingly rare when compared to *defectors,* the common type. If we assume that individuals interact at random with respect to behavioral strategy, the average fitness of both the common type *defectors* and rare mutant *reciprocators* will be dominated by interactions with the common type. *Reciprocators* do worse against *defectors* than *defectors* do against themselves. The reason is that *reciprocators* cooperate in the first round. Because they are mostly likely to interact with a *defector*, reciprocators receive "sucker's payoff." So, mutant *reciprocators* cannot invade a population of *defectors* and cooperation will not evolve.

Axelrod and Hamilton (1981) argued, however, that in many animals, there is some assortment to social interactions, perhaps due to low dispersal rates or kin-detection mechanisms. If so, those with the rare *reciprocator* mutation would be more likely to interact with other such mutants than chance alone would dictate. To capture this intuition, Axelrod and Hamilton introduced an exogenous assortment parameter, which could arise through relatedness, measuring the degree to which *tit for tat* types were likely to interact with other *tit for tat* types. With this assumption, Axelrod and Hamilton found a powerful synergy between kin selection and reciprocity, which work together to de-stabilize the uncooperative equilibrium, driving the evolution of cooperation through direct reciprocity. As can be seen in Figure 1, a similar synergy, for the same reason, exists with indirect reciprocity; even with low average relatedness, cooperation through indirect reciprocity readily invades an asocial population (Panchanathan and Boyd 2003).

[INSERT FIGURE 1 HERE]

To summarize, if we believe that reputations were widely broadcast through gossip and communities were long-lived during the course of human evolution, both of which plausibly fit the ethnographic record of small-scale societies, then the conditions would have been ripe for the evolution of cooperation, and so we might consider adding indirect reciprocity to our emerging explanation for human cooperation. However, the model of indirect reciprocity just presented only considers cooperation in the context of dyads. To be sure, dyads are ephemeral; hence the need for reputation to govern behavior. Still, our original goal was to explain the evolution of *group* cooperation among multiple unrelated individuals. Can the same logic that explains cooperation in the context of dyads be extended to larger groups?

4. Indirect reciprocity in large groups

To evaluate the hypothesis that reputation might explain the evolution of group cooperation, we elaborate Suzuki and Akiyama's (2007) model of indirect reciprocity in groups of arbitrary size. We begin as before with an infinite population. Individuals are randomly sampled into groups of size n (where $n \geq 2$). Members of each group engage in a one-shot public goods game, in which each individual can contribute to the collective, providing a benefit b that is shared equally among all group members, including himself, at a personal cost c. With a fixed probability w, individuals form new groups of size n and engage in another public goods game. The public goods game captures the idea of costly collective action that we are after: an individual's payoff is maximized by refusing to contribute, whereas the group's welfare increases with each contribution, assuming $b > c > b/n > 0$.

As discussed in the previous section, individuals know the reputation of any particular group member with a fixed probability q; thus, $1-q$ represents the fixed probability of not knowing the reputation of a particular individual (e.g., a total stranger), and the probability of knowing the reputations all of $n-1$ group members will be q^{n-1}. In this model, there is no inaccuracy in reputation knowledge, only ignorance. The community-wide reputation rule is such that individuals are "good" so long as they cooperate when *all* other group members are "good" or of unknown standing; refusing to contribute when there is at least one disreputable group member is permissible. If an individual falls into bad standing, good standing can be regained the next time that individual contributes to the public good, regardless of the group's composition in terms of strategies or reputations. We consider two behavioral strategies: *reciprocators* contribute to the public good so long as there are no known "bad" group members; *defectors* never contribute.

Assuming groups are large ($n >> 2$), the *reciprocator* strategy is evolutionarily stable if $qwb > c$. This is exactly the same condition as in the 2-person model (Condition (1)): cooperation is evolutionarily stable if the product of the probability of knowing the reputation of each social partner, the shadow of the future, and the benefit from the public good exceeds the personal cost of contribution. However, as discussed above for dyadic reciprocity, a thorough evolutionary explanation requires demonstrating that rare *reciprocators* can plausibly invade a population of *defectors*. As we noted, for dyadic reciprocity, direct or indirect, a little bit of exogenous assortment, such as kin-biased interactions, is sufficient to de-stabilize the uncooperative equilibrium, driving evolution to cooperation; but as groups get much larger than 2, the required assortment quickly grows to implausible levels (Figure 2).

[INSERT FIGURE 2 HERE]

Why does reciprocity work for small groups, but not for large ones? Dyadic reciprocity, whether direct or indirect, works because reciprocators can enjoy the benefits of cooperation, while excluding defectors. While the unfortunate reciprocator who happens to be paired with a cheat suffers, reciprocators as a whole do quite well for themselves. In groups much larger than a dyad, the situation is different (for a thorough treatment of the direct reciprocity case, see Boyd and Richerson 1988 or Sripada 2005). The only recourse open to a reciprocator who happens to find himself interacting with a cheat is to withhold cooperation (defect on the defector). Whereas a reciprocator can

withhold from single cheats in the dyadic case, refusing to contribute in a group situation penalizes other cooperators along with the cheat. Withholding help from those who do not help, the essence of reciprocity, becomes a blunt weapon in large groups, and thus reciprocity ceases to be a powerful driver of social evolution as group sizes increase.

We have modeled a situation in which group cooperation based on either direct or indirect reciprocity can be evolutionarily stable (Condition (1)), but invasion (the evolution of cooperation from an initially uncooperative state) becomes increasingly improbable as group size increases. When *reciprocators* are rare, most groups will consist of *defectors*. The only way for *reciprocators* to gain a foothold is if exogenous assortment is sufficiently high that *reciprocators* are grouped with *only* other *reciprocators*. When groups are dyads, a little bit of exogenous assortment will create a sufficient number of such dyads to prime the evolutionary pump; when groups are large, unless exogenous assortment is nearly complete, *reciprocators* will most likely find *at least* one *defector* in their midst, crashing cooperation (Figure 2).

To summarize, whereas indirect reciprocity might be a plausible model of cooperation enforced by reputation in a dyadic context, it is unlikely to be an explanation of large-scale cooperation. The power of reciprocity is in channeling cooperation to cooperators and withholding it from cheats. As group size increases, reciprocity by itself loses precision, being unable to channel cooperation to cooperators and withhold it from cheats. As we will see in the next section, however, there is a way of tapping the power of dyadic reciprocity to sanction defectors in collective action settings.

5. Social exclusion as targeted sanctioning

The provisioning of public goods is problematic because of the free-rider advantage. Reciprocity doesn't solve this problem when scaled up to larger groups. Effective punishment institutions, wherein the cost of being punished exceeds the cost of contributing, could solve the free-rider problem (Boyd and Richerson 1992), but punishment presents a second-order free-rider problem: being costly to administer, and yet contributing to collective welfare, punishers would seem to do worse than individuals who cooperate by contributing to first-order collective action but defect in the second-order punishment context. Although proposals have been made on how to solve this second-order problem of enforcing cooperation through punishment (Boyd and Richerson 1992; Gintis 2000; Henrich and Boyd 2001; Boyd et al. 2003), they generally do not involve reputation mechanisms and thus lie outside the scope of this paper. (A couple of exceptions, which involve signaling the ability or willingness to incur the costs of punishment, are discussed below in the section on signaling.)

One way that reputation-based mechanisms could stabilize group cooperation is by identifying free-riders and excluding them from the benefits of subsequent social exchange (Milinski et al. 2002; Panchanathan and Boyd 2004). The details of how this can be modeled are described in Box 2. The basic framework considers two types of cooperative interaction: a public goods game, and a mutual-aid game. Behavior in both contexts affects reputation, but defecting in the public goods game tarnishes reputation more than impermissibly defecting during mutual aid. The results of this analysis indicate that group cooperation can be evolutionarily stable when mutual aid is used as a carrot to induce public goods contribution (the benefits of social exchange are

withheld from non-contributors to the public good; Panchanathan and Boyd 2004). Like costly punishment (Boyd and Richerson 1992; Boyd et al. 2003), social exclusion (withdrawal of mutual aid) can induce group cooperation when the foregone benefits of mutual aid exceed the cost of contributing to the collective action. Unlike costly punishment, social exclusion does not suffer a corresponding second-order free rider problem, providing both a collective benefit (reducing free-riding) and a private one (not having to help those in need, specifically those with reputations as free-riders). As long as shunning free riders is socially permissible, the shunner does not suffer a reputation cost and can continue to enjoy the benefits of social exchange. There is no incentive to refuse to withhold help from free riders as this doesn't harm your own reputation, and you have no incentive to stay in the good graces of someone who will never aid you in your time of need.

[INSERT BOX 2 HERE]

Can this mechanism allow cooperation to evolve from an initially uncooperative ancestral state? As with the dyadic indirect reciprocity model presented in section 3, there is a powerful synergy between exogenous assortment and reciprocity in this combined collective-action plus mutual-aid model (Figure 3). Unlike the model of collective action with indirect reciprocity reviewed in section 4, the mutual aid model is not very sensitive to group size. When refusing to contribute to the public good is the only means to sanction free riders (the standard direct and indirect reciprocity cases), cooperation can gain a foothold when initially rare only if exogenous assortment is extremely high, sheltering mutant reciprocators in groups of like-minded reciprocators; just one defector can crash the party. In contrast, the collective action plus mutual aid model is not sensitive to group size because one defector does not have a big effect. To be sure, a defector will reap the benefits of free riding on the public good, but he is subsequently shunned from mutual aid. Because mutual aid decisions are dyadic—each potential helper decides whether or not to help a particular recipient—free riders can be singled out and excluded without jeopardizing ongoing social exchange between reciprocators, irrespective of the group size.

[INSERT FIGURE 3 HERE]

6. Signaling strategies and collective action

"Reputation" can have various meanings. The meaning we have so far employed is that individuals who play by the rules of reciprocity maintain a "good" reputation and enjoy the benefits of social exchange; those who do not play by the rules end up with "bad" reputations. An alternate way of conceptualizing reputation draws on signaling theory. This theory, with branches in both economics (Spence 2002) and biology (Johnstone 1997), analyzes certain kinds of traits as signals of underlying qualities that vary between individuals. For example, having large or brightly-colored tail feathers signals vigor and good health in many bird species; similarly, success as a warrior may signal likely prowess in other competitive arenas, while frequent unreciprocated donations of blood may signal both health and generosity (Lyle et al. 2009). In this sense, signaling strategies can be viewed as a means of establishing a reputation for ability or willingness to cooperate.

A key issue in any signaling system is how honesty is maintained; without this, there is no reason to expect observers to believe the signals they perceive. There are several possible answers, depending partially on context (Maynard Smith and Harper 2003; Cronk 2005). If the signaler and the observer have no conflicting interests, then there is no incentive for dishonesty. Complete coincidence of interests is expected to be rare, however, and this is almost certainly the case when we consider the evolution of group cooperation between unrelated individuals. In such cases, a more plausible guarantor of honesty involves *costly signaling*.

The fundamental requirements for stable costly signaling can be summarized as follows (Johnstone 1997; Bliege Bird and Smith 2005):

(1) individuals vary in one or more socially-relevant attributes ("quality") that are difficult to perceive directly (e.g., immune competence, cognitive abilities, social network size);

(2) signal costs or benefits are quality-dependent (e.g., lower quality signalers pay higher marginal signal costs);

(3) the best move for signal observers is to respond in ways that also benefit the signaler (e.g., forming alliances with high-quality signalers, choosing them as mates, or deferring to them in competitive contexts).

Note that these conditions apply to costly signaling in general, whether individuals signal their quality by "selfish" acts (e.g., signaling access to wealth via extravagant consumption of luxury goods) or cooperative ones (e.g., signaling fighting ability by defending the village).

To see how signaling might motivate costly collective action, one can add a signaling dynamic to a standard public goods (n-player Prisoner's Dilemma) interaction (Gintis et al. 2001). Suppose that providing some collective benefit at personal cost c constitutes an honest signal of high quality—for example, of productive efficiency, wealth, or social network size. This signal induces at least one observer (who may or may not be involved in the public goods game, and hence need not receive a share of the collective good) to interact with the signaler in a mutually beneficial manner (as noted in condition 3, above). Such a response boosts the signaler's expected payoff by some amount s. If $s > c$, even unilateral cooperation (the sucker's payoff in a standard public-goods game) can be evolutionarily stable (Gintis et al. 2001). Put simply, signaling dissolves the collective action problem as long as the signaler's gain from signaling exceeds her cost in producing the collective good. For example, a Kwakiutl chief who hosts a potlatch and magnanimously provides gifts to dozens of guests could gain status and other benefits (perhaps at the expense of a rival chief) that in the long run offer fitness gains exceeding his potlatching costs (Boone 2000).

For the logic of this model to hold, condition 3 (above) must be met: there must be a personal benefit to granting status to those who provide collective benefits. In many cases, this is plausible: high contributions to collective action may reliably signal qualities that make one a preferred ally or mate (e.g., health, vigor, ability to generate wealth, a strong social network). In evolutionary terms, status itself is but a proximate mechanism to gain access to fitness benefits. If granting status does not yield the grantor some benefit, such as preferential access to the high-status individual , then there will be a temptation to free-ride on the status-granting actions of others (Smith and Bliege Bird 2000).

An interesting aspect of the costly-signaling approach is that signaling doesn't just permit the evolution of collective action where it might otherwise be problematic, it may actually feed off it. This is because signaling by providing collective goods increases the signal's "broadcast efficiency" (Smith and Bliege Bird 2000): sharing food at a feast attracts a larger audience than sharing the same amount of food with your neighbor, and fighting valiantly to defend your village broadcasts your quality to many more people than fighting in a bar-room brawl. Thus, for appropriate settings and signals, the potential for signaling to solve collective action problems would seem to increase with group size—the opposite of the usual case.

Signaling ability versus intent

There are two distinct forms of signaling that are relevant to the evolution of cooperation. First, providing public goods without direct compensation may signal that the provider has underlying qualities that make him or her a desirable ally or mate, or a formidable competitor (Smith and Bliege Bird 2000). A successful turtle hunter among the Meriam of Torres Strait helps feed dozens or hundreds of fellow islanders at communal feasts, and receives no direct material reward; but his status rises, and with it his reproductive success (Smith et al. 2003). Indeed, the sharing of large game is a ubiquitous feature of many societies, and the fact that more successful hunters gain in status and reproductive success can be interpreted in terms of signaling theory (Smith 2004). A similar dynamic may be at work in the voluntary provisioning of a broad array of public goods, from sponsoring feasts and ceremonies (Boone 1998) and financing the construction of monuments (Neiman 1998) to leading war parties (Patton 2005). Similarly, if individuals vary in some measure of quality (such as competitive ability or social network size) that makes it less costly for them to monitor and punish those who fail to contribute to collective action, then the signal of higher quality can be the collective good of enforcing group cooperation (Gintis et al. 2001). If audience members preferentially chose such individuals as allies or mates, or defer to them in other contexts besides collective action, such group-beneficial signaling strategies may be evolutionarily successful.

These arguments and examples exemplify "conspicuous expenditure" (Veblen 1898) in which the signaler demonstrates the ability to absorb unusually high costs due to underlying qualities such as good health, vigor, productivity, wealth, or strong social networks (Bliege Bird and Smith 2005). An alternative set of arguments about how signaling strategies may drive the evolution of cooperation focus on signals of prosociality or cooperative intent. If you invite your neighbor or friend to dinner at your home, it is unlikely to impress as a signal of extraordinary wealth or productivity, but it does signal some degree of commitment to an ongoing relationship (as do such trivial expenditures of time and wealth as greeting cards, or attendance at ritual events such as weddings). Just such a dynamic seems to be at work in the *hxaro* gift-exchange system of the Ju/hoansi (Wiessner 2002), in which individuals cultivate a network of often far-flung allies who can be relied on for hospitality or assistance when needed.

But signaling commitment to a dyadic relationship is not the same as signaling prosociality or commitment to the collective good. As an example of the latter, Gurven and colleagues (2000) have shown that among the Ache, generous hunters are more likely to be aided by fellow villagers when they are incapacitated by illness or injury, while stingy ones are penalized; this effect is independent of hunting productivity, since

poor hunters who share their meager harvest generously receive just as much aid as more productive generous hunters, whereas productive hunters who share a smaller proportion of their harvest receive no more aid than non-producers (Gurven et al. 2000). This example demonstrates how fuzzy the boundary between signaling explanations and indirect reciprocity explanations can be (see Bergmuller et al [2007] for a useful categorization of these different mechanisms). Generosity in sharing products of the hunt is a signal of both hunting skill and cooperative intent, and to the extent that the hunter shares widely and without direct reciprocation it can be seen as providing a public good. In turn, those who aid the generous hunter when he is incapacitated are participating in the mutual-aid phase of the linked games outlined above in section 5. Similar arguments have been made for analyses of contribution to collective goods in other contexts, including horticultural societies (Price 2003) and industrialized urban settings (Barclay and Willer 2007; Bereczkei et al. 2007; Nelissen 2008). A recent model shows how a costly signal of commitment to punish non-cooperators if there are sufficient other punisher-signalers can help stabilize group cooperation (Boyd et al. 2010).

Serious objections can be raised to these "signaling of intent" arguments, however. What is to stop someone from signaling an intent to cooperate, receiving a benefit, and then violating the trust by refusing to cooperate? Although some have argued that "subjective commitment" can serve as a guarantor of future cooperation (Frank 1988; Hirshleifer 1987; various authors in Nesse 2001), this has been criticized for lack of solid game-theoretical basis (Adams 2001; Ross and Dumouchel 2004). However, such objections are not necessarily fatal. First, signals with sufficient initial cost (such as an engagement ring) may ensure that the signaler is committed to a lengthy period of cooperative interactions in which this signal cost can be recouped (Carmichael and McLeod 1997; Smith and Bliege Bird 2005; Bergstrom et al. 2008). For example, participation in religious ritual can serve as a costly signal of one's commitment to provide collective goods to one's group under circumstances in which those seeking only short-term gains (i.e., free riders) would be tempted to defect (Irons 2001; Sosis and Alcorta 2003; Sosis and Bressler 2003). The costs of ritual observances include time (e.g., attending services, days of rest), resources (e.g., financing communal rituals, observing taboos), and sometimes even morbidity or mortality risks (e.g., tests of faith, ritual mutilation). Signaling theory prompts the hypothesis that, by paying these costs, individuals signal to others that they are indeed committed to long-term collective action in their social group. Signal costs help secure this commitment if they can only be recouped through group membership over the long run.

Second, even if a signal of cooperative intent is not secured with an initial or endogenous cost, breaking this commitment might be so socially costly as to deter most individuals from practicing such deception. It is arguably much more harmful to one's reputation to defect after promising cooperation—to defect on one's allies—than to defect without violating any pre-existing signal of commitment (Shinada et al. 2004). Just why such hypocrisy elicits greater condemnation than simple defection is an interesting evolutionary problem in itself. In any case, much further research needs to be done on the theory and empirics of signals of cooperative intent.

7. Equilibrium selection and the evolution of norms

Explanations for group cooperation which involve enforcing cooperation through imposing some cost, be it punishment or social exclusion, have one common feature: the cost imposed on defectors must exceed the cost of contributing to the collective good. Similarly, explanations of collective action as signaling strategies require that signal costs be less than the gains signalers receive from influencing observers. Note, however, that the benefit of collective goods contribution (e.g., B in the collective action and mutual aid model detailed in Box 2) does not enter into the stability conditions for either indirect reciprocity or costly signaling. Because the public good flows to all group members equally, whether or not they contributed, only the costs of contributing matter. This means that the same mechanism that stabilizes group cooperation, whether through indirect reciprocity or signaling, can stabilize any social norm, including norms prescribing welfare-neutral or even welfare-decreasing behaviors (where welfare refers to average benefit to members of the collective—mean fitness, if you like).

This counter-intuitive result is a general feature of models that explain group cooperation through threats of punishment or social exclusion (Boyd and Richerson 1992; Henrich and Boyd 2001; Panchanathan and Boyd 2004). If the threat of punishment or exclusion from social exchange is sufficiently high, it pays to follow the social rules, even if the rules lead to maladaptive outcomes. A similar logic applies to signaling arguments. Selection favors signaling when it pays off for the signaler, even when the signaling equilibrium makes everyone worse off (Bergstrom and Lachmann, 1997), as with some cases of conspicuous consumption as signals of wealth, or beating up innocent bystanders as a signal of vigor.

To illustrate this point, let's revisit the model of large-scale cooperation and mutual aid from section 5. We begin with a large population of individuals, sampled into groups of size n, playing a one-shot public goods game, and then a series of mutual aid games. To make this example more concrete, let's suppose the public good involves forest clearing to make room for the cultivation of annuals. Individuals who do not chop down trees are excluded from subsequent mutual aid, which might be something like the community helping one member build his house. Now, imagine a small group of progressives who propose that forest clearing is, in fact, not a good idea; instead, they advocate conserving forest resources. For the sake of argument, let us suppose that conservation results in higher mean welfare than forest clearing. Will selection favor the adoption of the new and improved social norm?

It turns out that both forest clearing and forest conservation are evolutionarily stable equilibria (for a formal derivation, see the supplementary information in Panchanathan and Boyd 2004). Why is an equilibrium with lower mean fitness stable against invasion by a strategy that could make the population better off? As with previous models, we need to consider how much exogenous assortment is needed for selection to favor the evolution of the new, welfare-enhancing social norm. It turns out that, unlike previous models, there is no synergy between exogenous assortment and reciprocity (Figure 4). When transitioning from one cooperative equilibrium to another, even when the new equilibrium is welfare-enhancing, exogenous assortment and reciprocity oppose one another. The more that social exchange matters, the more exogenous assortment needed to destabilize the entrenched equilibrium and drive social evolution to the new equilibrium.

[INSERT FIGURE 4 HERE]

To understand why this is so, let's begin by assuming no exogenous assortment: groups are formed at random with respect to behavioral strategy. If we want to know under what conditions selection favors the transition from forest clearing to forest conservation, we must assume that, from the outset, most of the individuals are forest clearers. Dismayed by the behaviors of the majority, the only recourse for the intrepid conservationist is to refuse to help the forest clearers in mutual aid. In the eyes of the majority, however, this behavior prompts retribution; to forest clearers, the conservationist is a defector, and thus labeled "bad" and excluded from mutual aid. Because we must assume that mutual aid looms large—if it didn't, social exclusion wouldn't be a meaningful threat and cheating would triumph—the rare conservationist suffers more from being shunned by the majority forest clearers than the majority clearers suffer from being shunned by the conservationist, even though the conservationist advocates the better collective action norm. As a result, the norm of conservation cannot spread under standard evolutionary dynamics, whether genetic or cultural. Adding exogenous assortment, meaning forest clearers are more likely to be in groups with other clearers and conservationists are more likely to be with other conservationists, rather than randomly-formed communities, helps only to a degree (Figure 4). When mutual aid is a powerful force, extremely high levels of exogenous assortment are required before selection favors the new social norm. It is plausible that this dynamic is relevant to many troubling cases of cultural conservatism, such as female genital cutting norms (Mackie 1996). Axelrod and Hamilton (1981) anticipated this kind of a result in their original formulation of direct reciprocity. They noted that assortment through a mechanism like kinship interacts with reciprocity to drive the evolution of cooperation—*tit for tat* invades a population of *defectors*. That assortment through kinship doesn't help *defectors* invade a population of *tit for tat* prompted Axelrod and Hamilton to write, "the gear wheels of social evolution have a ratchet." This ratchet, however, only works when considering the evolution of cooperation from an initially asocial population; when considering the transition from one cooperative equilibrium to another, reciprocity and kinship operate antagonistically.

The reason why exogenous assortment and reciprocity operate synergistically in previous models of reciprocity, while they oppose each other in this model, lies with what the entrenched majority is doing (i.e., the ancestral state of the population). In the models of reciprocity considered above, and Axelrod and Hamilton's model of direct reciprocity, the ancestral condition is uncooperative, based on the *defector* strategy. As we introduce exogenous assortment, reciprocators are more likely to meet one another, bounding cooperation within reciprocating dyads, while defectors are left to interact with one another, gaining nothing from these interactions. The more exogenous assortment, the more reciprocators keep the benefits of social exchange to themselves, and the less some of it bleeds out to defectors. When considering the evolutionary transition between two different cooperative equilibria based on reciprocity, we must keep in mind that the ancestral condition is a cooperative world in which one particular norm is prescribed, while the mutants offer a different vision of what constitutes cooperation. Although introducing exogenous assortment helps the rare mutant do a little better during the collective action phase, the common types, unlike defectors in the previous models, are actively engaged in mutual aid with one another. When reciprocity

looms large, being a part of social exchange is paramount, and so it pays to do what the majority wants.

This result tells us that, in cases where collective action is stabilized by the threat of punishment or social exclusion from reciprocity, or facilitated by a link to signaling strategies, selection amongst different stable equilibria could be an important force in shaping social evolution. While the literature on equilibrium selection is too large to review here (see Young 1993; Samuelson 1997; Bowles 2004; Boyd and Richerson 1990, 2002), the take home message is that mechanisms that stabilize collective action norms via reciprocity or signaling don't necessarily favor welfare-enhancing norms. Furthermore, indirect reciprocity tends to homogenize social groups with respect to norms; when reputations for upholding norms matter, diversity within communities is not tolerated. The variance that remains will be distributed between groups. A similar logic might apply to signaling, as different social groups may have their own idiosyncratic pathways to gain status through signaling. As variation is the fuel of selection, more between-group variation means more potential for between-group selection.

8. Conclusions

As research on the evolution of cooperation marches forward, the concepts and explanations multiply. While it's obvious that kin selection and direct reciprocity are insufficient to explain the extent of group cooperation in humans, it's not yet clear which other mechanisms should be admitted and how to distribute explanatory responsibility among them. In our estimation, a complete account will likely include cultural transmission, reputation, punishment, signaling, social institutions, and multi-level selection. In this paper, we have reviewed some of the recent work on reputation, attempting to clarify how the concept has been used both in the context of reciprocity and signaling.

In section 3, we showed how reputation-based reciprocity (termed indirect reciprocity) can explain dyadic cooperation. When individuals can effectively track the goings on of their neighbors and communities persist, reciprocity based on reputation, in which reciprocators channel cooperation to those who have reputations for following the rules of reciprocity, can evolve and thrive. In section 4, we extended the basic indirect reciprocity model to allow for larger groups. While the stability condition for group cooperation through indirect reciprocity is identical with the dyadic case, the evolvability condition differs markedly. As group size increases, the exogenous assortment needed to get reciprocity off the ground rapidly increases. As in the case of direct reciprocity, simply scaling up indirect reciprocity to large groups won't do the trick. In section 5, we showed how linking a collective action game to a mutual aid game (really just a series of dyadic indirect reciprocity games) can offer an explanation for group cooperation. Linking collective action to dyadic indirect reciprocity can focus the sanctioning power of social exclusion (i.e., not being included in reciprocity) on collective action cheats, and thus offers an attractive explanatory mechanism for group cooperation.

In section 6, we reviewed models of signaling and group cooperation. Unlike reciprocity, in which reputation means playing by the rules of social exchange, in the case of signaling, reputation links behavior to the value of latent variables. Individuals signal their quality through costly action, like throwing elaborate and wasteful feasts or

provisioning others with big game. Attending to the signals benefits recipients by allowing them to preferentially interact with high quality partners, which would be useful when seeking coalition partners or mates. So long as signal costs are state-dependent and the benefits of signaling outweigh the costs, costly signaling can be a force in generating group cooperation. The other, more controversial, way in which signaling might relate to group cooperation is with signals of intent. Here, there need not be a state-dependent cost to signaling. Instead, signals are devices that reveal the intentions of the signaler. Such systems could be kept honest in at least three ways: signaling comes along with a internally-enforced commitment, signals have costs that can only be recouped upon trustworthy behavior, or hypocrites are extraordinarily punished.

In section 7, we discussed equilibrium selection. In many models of group cooperation, whether based on punishment, indirect reciprocity, or signaling, the stability condition for cooperation doesn't include the benefit of cooperation. This counter-intuitive result means that the mechanisms we are interested in are no more likely to generate group cooperation than welfare-neutral or even welfare-decreasing traits. In cases in which multiple equilibria are possible, mechanisms which select between equilibria become important.

Box 1. Glossary

Assortment: We consider two types of assortment. Reputation and signaling endogenously generate behavioral assortment (e.g., reciprocators only cooperate with those with a good reputation). When modeling the invasion of reciprocating strategies, we introduce exogenous assortment (e.g., kin-biased interactions), which result in reciprocators interacting with other reciprocators in a non-random fashion.

Collective action problem: Any situation where several or many individuals must cooperate in order to produce some collective good.

Collective good: Any good or service provided to the members of some group (coalition, village, organization, nation, etc.) through the efforts of some or all of its members; similar to a *public good*, which has a more restrictive meaning.

Costly signaling: Providing information about a hidden attribute by producing a signal that is too costly for those lacking the attribute to profitably incur.

Direct reciprocity: Individual A helps individual B if B had previously helped A.

Evolutionarily stable strategy: A behavioral strategy which, when adopted by the majority, cannot be invaded by rare, mutant strategies.

Free rider: One who gains the benefits of cooperation (e.g., consumes a portion of a *collective good*) without paying the costs (e.g., not contributing to production of the good).

Indirect reciprocity: Individual A helps B if B had previously helped C.

Non-rival: A good or service whose consumption by some does not reduce the amount available to others—for example, lighthouse beacons or TV broadcasts.

Non-excludable: A good or service available to all group members, regardless of their contribution to providing the good.

One-shot game: A game structure in which the same individuals are not paired or grouped together to repeat the game; contrast with *repeated game*.

Prisoner's dilemma: A situation in which an individual has one of two options, cooperate or defect, and does best by defecting regardless of his opponent's behavior.

Public good: A subset (although sometimes used as a synonym) of *collective goods*. Pure public goods are *non-rival* as well as *non-excludable*, whereas collective goods can be either rival or non-rival, but are non-excludable to the group of interest, and thus include both pure public goods and common-property resources (as defined by Ostrom 1990 and others).

Public goods game: A game in which individuals can contribute to providing a *public good* (or, less stringently, a *collective good*) at some personal cost.

Repeated game: A game in which the same set of individuals interact with each other repeatedly (and thus have ongoing opportunities for reciprocity or sanctioning).

Second-order problem: When a solution to one collective action problem introduces a new collective action problem. For example, punishing *free-riders* can enforce cooperation, but the second-order problem is that there is now a temptation to free ride on the enforcement efforts of others.

Tit for tat: A proposed solution to direct reciprocity games, in which the strategy is to cooperate on the first round, then copy the other player's move in each subsequent round (defect if they defect, cooperate if they cooperate).

Box 2. Linking Indirect Reciprocity to Collective Action

How can the reputation-based mechanisms of indirect reciprocity be used to stabilize collective action? Inspired by the experimental results of Milinski et al. (2002), Panchanathan and Boyd (2004) proposed one model of how this might work. Begin with an infinite population. Individuals are randomly sampled into groups of size n (where $n \geq 2$). Within each group, individuals participate in a one-shot public goods game: each individual can contribute to the collective, providing a benefit B that is shared equally among all group members, including himself, at a personal cost C, assuming $B > C > B/n > 0$ (we use capital letters to distinguish these benefits and costs from those of mutual aid, shown below). Group members then play a repeated mutual aid game (Sugden 1986). In each round, one randomly-selected group member is designated as needy. The $n-1$ other group members can each provide a benefit b to the needy recipient, at a personal cost c. Each of the $n-1$ potential helpers makes an individual decision whether or not to help the needy recipient. With a fixed probability w, the individuals within the group engage in another round of mutual aid. Again, one of the group members is randomly designated as the needy recipient, while the $n-1$ group members can help the recipient (following Panchanathan and Boyd 2004).

Each individual has a reputation that summarizes her previous behavior in both collective-action and mutual-aid interactions. For simplicity, and without loss of generality, we assume that each group member knows the reputations of all other group members (i.e., $q=1$). The reputation rule we consider has the following features: (1) individuals who refuse to contribute to the public good fall into disrepute and can never redeem themselves (the results of this model do not turn on the severity of this assumption; all that is required is that withholding from the collective action has more severe reputation consequences than refusing to help a worthy recipient during a bout of mutual aid); (2) those who contribute to the public good enter the mutual aid phase in good standing; (3) failure to help a "good-standing" recipient during mutual aid results in "bad" standing; and (4) good standing can be regained by subsequently helping a recipient during a bout of mutual aid. We consider two behavioral strategies: *shunners* contribute to collective action and help good-standing recipients during mutual aid; *defectors* neither contribute to the public good nor provide aid to needy individuals.

Assuming that groups are large ($n >> 2$), the stability condition for *shunners* is approximately:

$$(b-c)/(1-w)>C \qquad \text{(B1.1)}$$

The term to the right of the inequality, C, represents the cost of contributing to the collective action. The term to the left of the inequality, $(b-c)/(1-w)$, represents the long-run benefit of participating in the mutual aid system. As w represents the fixed probability of an additional round of mutual aid, $1/(1-w)$ represents the expected number of mutual aid rounds. $b-c$ represents the net benefit an individual can expect on a per round basis by participating in mutual aid. The participant will be needy about once every n rounds. When this happens, he will receive the benefit b from the $n-1$ other participants. When he is not needy, which happens with a probability $1-1/n = (n-1)/n$, the participant will have to help one of his neighbors, paying the cost c. The number of rounds of mutual aid, $1/(1-w)$, multiplied by the net benefit of one round of

social exchange, $b-c$, results in the net benefit of mutual aid. *Shunners*, because they contribute to the collective action, enjoy this benefit; *defectors*, because they do not, forgo this benefit. When the benefit of mutual aid exceeds the cost of collective action contribution, *shunner* is an evolutionarily stable strategy and cooperation through reputation-based reciprocity thrives. Seen another way, the inequality states that group cooperation is stabilized by indirect reciprocity when the sanction (here, the withdrawal of mutual aid, $(b-c)/(1-w)$) exceeds the benefit of free riding (here, the cost of contribution to the collective action, C). This is a common feature of models that explain cooperation through punishment: cooperation is evolutionarily stable (or individually self-interested) when the cost of being punished exceeds the cost of cooperating. As in other reciprocity models, the uncooperative equilibrium, a population of *defectors*, is also evolutionarily stable.

References Cited

Adams, Eldridge S. (2001) Threat displays in animal communication: handicaps, reputations, and commitments. In *Evolution and the capacity for commitment*, ed. Randolph M. Nesse, pp. 99-119. NY: Russell Sage Foundation.

Alexander, Richard D. (1987) *The Biology of Moral Systems*. Hawthorne, NY: Aldine de Gruyter.

Axelrod, Robert (1984) *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert and William D. Hamilton (1981) The evolution of cooperation. *Science* 211:1390-1396.

Barclay, Pat and Robb Willer (2007) Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London, Series B* 274:749-753.

Bereczkei, Tamas, B. Birkas, and Z. Kerekes (2007) Public charity offer as a proximate factor of evolved reputation-building strategy: an experimental analysis of a real-life situation. *Evolution and Human Behavior* 28(4):277-284.

Bergmuller, R., R.A. Johnstone, A.F. Russell, and R. Bshary (2007) Integrating cooperative breeding into theoretical concepts of cooperation. *Behavioural Processes* 76: 61–72.

Bergstrom, Carl T., Ben Kerr, and Michael Lachmann (2008) Building trust by wasting time. In *Moral Markets: The Critical Role of Values in the Economy*, ed. P. Zak, pp. 142-156. Princeton University Press.

Bergstrom, Carl T., and Michael Lachmann (1997) Signalling among relatives. I. When is signalling too costly? *Philosophical Transactions of the Royal Society of London, Series B*, 352:609-617.

Bliege Bird, Rebecca L. and Eric Alden Smith (2005) Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology* 46(2):221-248.

Boone, James L. (1998) The evolution of magnanimity: when is it better to give than to receive? *Human Nature* 9(1):1-21.

Boone, James L. (2000) Status signaling, social power, and lineage survival. In *Hierarchies in action: cui bono?* ed. Michael W. Diehl, pp. 84-110. Carbondale, IL: Center for Archaeological Investigations, Southern Illinois University.

Bowles, Samuel (2004) *Microeconomics: Behavior, Institutions, and Evolution*. Princeton University Press, Princeton.

Boyd, Rob, Herbert Gintis, and Samuel Bowles (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328(5978): 617-620.

Boyd, Rob, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences USA* 100(6):3531-3535.

Boyd, Robert and Peter J. Richerson (1988) The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology* 132:337-356.

Boyd, Robert and Peter J. Richerson (1990) Group selection among alternative evolutionary stable strategies. *Journal of Theoretical Biology* 145: 331-342.

Boyd, Robert and Peter J. Richerson (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13:171-195.

Boyd, Robert and Peter J. Richerson (2002) Group beneficial norms can spread rapidly in a structured population. *Journal of Theoretical Biology* 215: 287-296.

Carmichael, H. L. and W. B. MacLeod (1997) Gift giving and the evolution of cooperation. *International Economic Review* 38:485-509.
Cronk, Lee (2005) The application of animal signaling theory to human phenomena: Some thoughts and clarifications. *Social Science Information* 44(4):603-620.
Darwin, Charles (1859) *The Origin of Species*. London: John Murray.
Darwin, Charles (1871) *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.
Fehr, Ernst and Simon Gächter (2002) Altruistic punishment in humans. *Nature* 415, Jan. 10:137-140.
Frank, Robert H. (1988) *Passions Within Reason*. NY: Norton.
Gintis, Herbert (2000) Strong reciprocity and human sociality. *J. of Theoretical Biology* 206:169-179.
Gintis, Herbert, Eric Alden Smith, and Samuel L. Bowles (2001) Cooperation and costly signaling. *J. of Theoretical Biology* 213:103-119.
Grafen, Alan (1984) Natural selection, kin selection and group selection. In *Behavioural Ecology: An Evolutionary Approach*, ed. J.R. Krebs and N.B. Davies, pp. 62-84. Sunderland, MA: Sinauer Associates.
Greif, A. (1989) Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *The Journal of Economic History* 49(4): 857-882.
Gurven, Michael, et al. (2000) "It's a wonderful life": signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior* 21(4):263-282.
Hamilton, William D. (1964) The genetical evolution of social behaviour, I, II. *J. of Theoretical Biology* 7:1-52.
Hammerstein, Peter (2003) Why is reciprocity so rare in social animals? A Protestant appeal. In *The genetic and cultural evolution of cooperation*, ed. Peter Hammerstein, pp. 83-93. Cambridge, MA: MIT Press.
Hardin, Garrett (1968) The tragedy of the commons. *Science* 162:1243-48.
Henrich, Joseph and Robert Boyd (2001) Why people punish defectors: weak conformist transmission can stabilize costly enforcement of between-group differences. *J. of Theoretical Biology* 208:79-89.
Hirshleifer, Jack (1987) On the emotions as guarantors of threats and promises. In *The Latest on the Best: Essays on Evolution and Optimality*, ed. J. Dupré, pp. 307-26. Cambridge: MIT Press.
Irons, William G. (2001) Religion as a hard-to-fake sign of commitment. In *Evolution and the capacity for commitment*, ed. Randolph M. Nesse, pp. 292-309. NY: Russell Sage Foundation.
Johnstone, Rufus A. (1997) The evolution of animal signals. In *Behavioural ecology: an evolutionary approach*, ed. John R. Krebs and Nicholas B. Davies, pp. 155-178. Oxford: Blackwell.
Lyle III, Henry F., Eric A. Smith, and Roger J. Sullivan. 2009. Blood donations as costly signals of donor quality. *Journal of Evolutionary Psychology* 7(4):1-24.
Mackie, Gerry (1996) Ending footbinding and infibulation: A convention account. *American Sociological Review* 61(6): 999-1017.
Maynard Smith, John (1964) Group selection and kin selection. *Nature* 201:1145-47.
Maynard Smith, John (1982) *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Maynard Smith, John and David Harper (2003) *Animal signals*. Oxford: Oxford U. Press.
Milinski, M., D. Semman, and H.J. Krambeck (2002) Reputation helps solves the 'tragedy of the commons.' *Nature* 415: 424–426.
Neiman, Fraser D. (1997) Conspicuous consumption as wasteful advertising: a Darwinian perspective on spatial patterns in Classic Maya terminal monument dates. In *Rediscovering Darwin: Evolutionary theory and archeological explanation*, ed. C. Michael Barton and Geoffrey A. Clark, pp. 267-290. Washington, D.C.: Archeological papers of the American Anthropological Association, No. 7.
Nelissen, Rob A. (2008) The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior* 29(4):242-248.
Nesse, Randolph M., ed. (2001) *Evolution and the capacity for commitment*. NY: Russell Sage Foundation.
Nowak, Martin A. and Karl Sigmund (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393:573-577.
Ohtsuki, Hisashi and Yoh Iwasa (2006) The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* 239(4):435-444.
Olson, Mancur (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.
Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
Panchanathan, Karthik and Rob Boyd (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. of Theoretical Biology* 224:115-126.
Panchanathan, Karthik and Rob Boyd (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432:499-502.
Patton, John Q. (2005) Meat sharing for coalitional support. *Evolution and Human Behavior* 26(2):137-157.
Price, Michael E. (2003) Pro-community altruism and social status in a Shuar village. *Human Nature* 14(2):191-208.
Ross, Don and Paul Dumouchel (2004) Emotions as strategic signals. *Rationality and Society* 16(3):251-286.
Samuelson, Larry (1997) *Evolutionary games and equilibrium selection*. Cambridge, MA: MIT Press.
Schaffer, William F. 1978. A note on the theory of reciprocal altruism. *American Naturalist* 112:250-254.
Shinada, Mizuho, Toshio Yamagishi, and Yu Ohmura (2004) False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. *Evolution and Human Behavior* 25(6):379-393
Smith, Eric Alden (2004) Why do good hunters have higher reproductive success? *Human Nature* 15(4):343-364.
Smith, Eric Alden and Rebecca L. Bliege Bird (2000) Turtle hunting and tombstone opening: Public generosity as costly signaling. *Evolution and Human Behavior* 21(4):245-261.

Smith, Eric Alden and Rebecca L. Bliege Bird (2005) Costly signaling and cooperative behavior. In *Moral sentiments and material interests: On the foundations of cooperation in economic life*, ed. H. Gintis, et al., pp. 115-148. Cambridge, MA: MIT Press.

Smith, Eric Alden, Rebecca L. Bliege Bird, and Douglas W. Bird (2003) The benefits of costly signaling: Meriam turtle hunters. *Behavioral Ecology* 14(1):116-126.

Sosis, Richard and Candace Alcorta (2003) Signaling, solidarity, and the sacred: the evolution of religious behavior. *Evolutionary Anthropology* 12(6):264-274.

Sosis, Richard and E. Bressler (2003) Cooperation and commune longevity: a test of the costly signaling theory of religion. *Cross-Cultural Research* 37:211-239.

Spence, Michael (2002) Signaling in retrospect and the informational structure of markets. *American Economic Review* 92(3):434-459.

Sripada, Chandra S. (2005) Punishment and the strategic structure of moral systems. *Biology and Philosophy* 20(4):767-789.

Sugden, R. (1986) *The Economics of Rights, Co-operation and Welfare*. Oxford: Basil Blackwell.

Suzuki, Shinsuke and Eizo Akiyama (2007) Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. *J. of Theoretical Biology* 245(3):539-552.

Taylor, M. (1976) *Anarchy and Cooperation*. Johny Wiley & Sons.

Trivers, Robert L. (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* 46:35-57.

Veblen, Thorstein (1898) *The theory of the leisure class*. NY: Macmillan.

Wiessner, Polly (2002) Hunting, healing, and *hxaro* exchange: a long term perspective on !Kung (Ju/'hoansi) large-game hunting. *Evolution and Human Behavior* 23(6):407-436.

Williams, George C., and Doris C. Williams (1957) Natural selection of individually harmful social adaptations among sibs with special reference to social insects. *Evolution* 11(1):32-39.

Young, H. Peyton (1993) The evolution of conventions. *Econometrica* 61(1): 57-84.

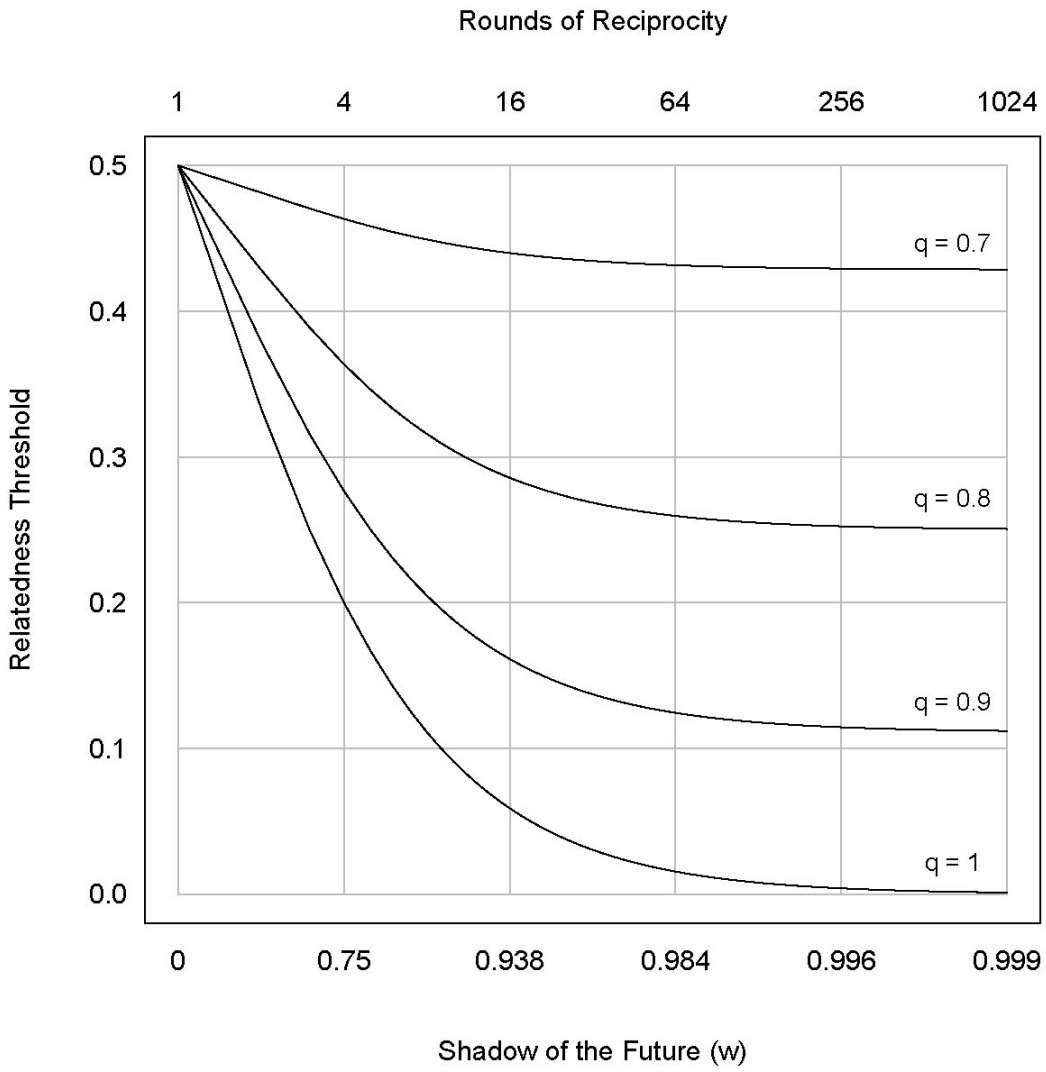

Figure 1. The minimum relatedness necessary for the *reciprocator* strategy to invade a population of *defectors* in a 2-person indirect reciprocity model. The strength of reciprocity, w, Axelrod's (1984) "shadow of the future," is depicted on the bottom x-axis on a log scale. The top x-axis shows the corresponding number of rounds of social exchange (e.g., w=0.75 implies 4 expected rounds). The cost of helping, c, is set to 1, and the benefit of receiving help, b, is set to 2. A b/c ratio of 2 represents the minimum requirement for altruism between full siblings according to Hamilton's rule. The curves represent the required relatedness as a function of the strength of reciprocity for different degrees of reputation quality (q = 0.7, 0.8, 0.9, 1.0). When the shadow of the future is sufficiently small, the minimum relatedness is near 0.5; for cooperation to evolve, interactions would have to be between full siblings. When reputation quality is high ($q \approx 1$) and the future casts a long shadow ($w \approx 1$), the minimum relatedness diminishes to low levels. As reputation quality decreases, despite many rounds of social interacion, the minimum relatedness for cooperation to evolve is high. When reputation quality is perfect (q=1), the minimum relatedness is the same as in Axelrod and Hamilton's (1981) model of direct reciprocity. There is a synergy between relatedness and reciprocity: as strength of reciprocity increases, the relatedness necessary for cooperation to evolve decreases, though it diminishes as reputation quality decreases.

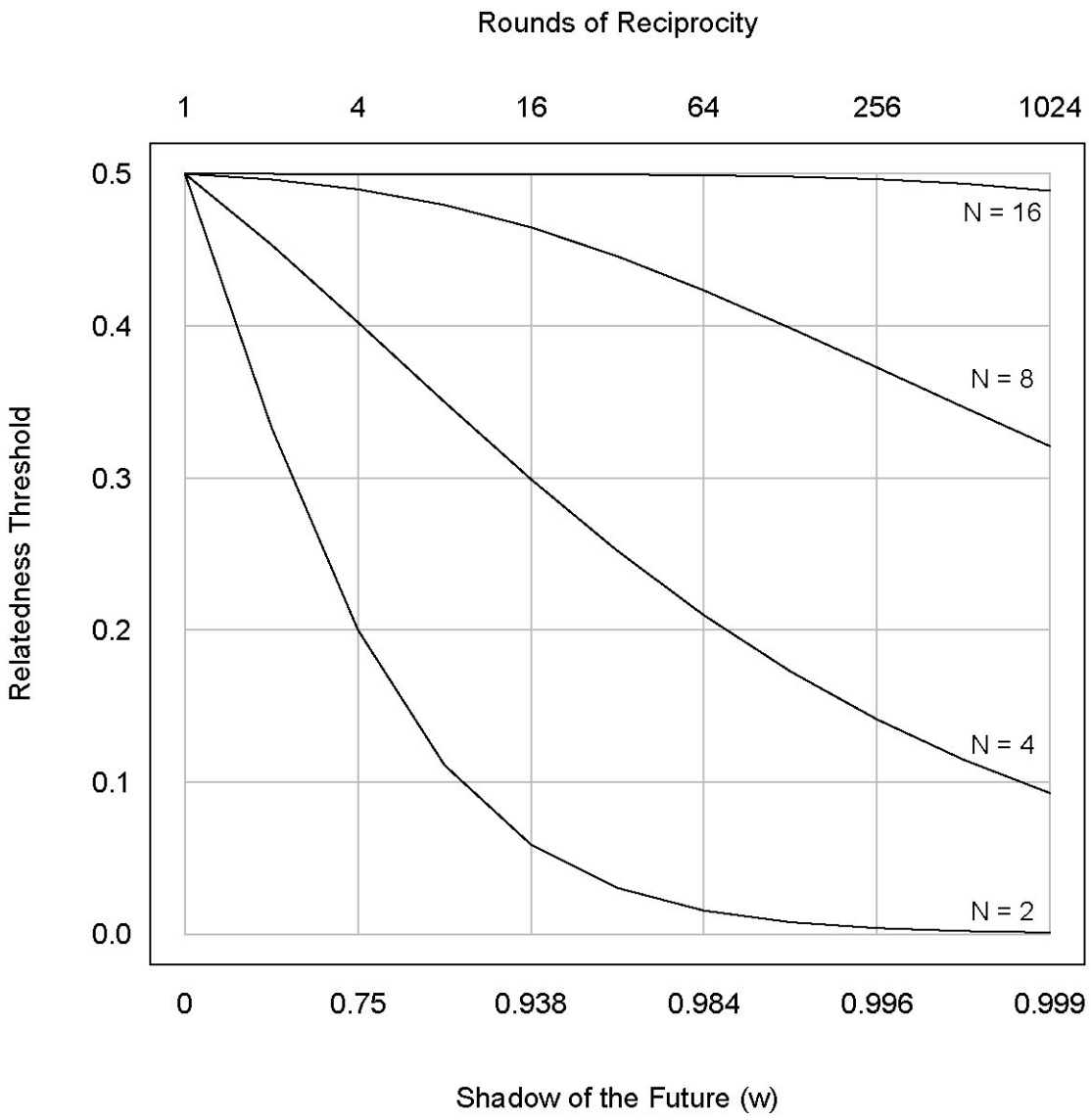

Figure 2. The minimum relatedness necessary for the *reciprocator* strategy to invade a population of *defectors* in an *n*-person indirect reciprocity model. The strength of reciprocity, *w*, Axelrod's (1984) "shadow of the future," is depicted on the bottom *x*-axis on a log scale. The top *x*-axis shows the corresponding number of rounds of social exchange (e.g., w=0.75 implies 4 expected rounds). The reputation quality (q) is set to 1, meaning that individuals can accurately track the reputation of everyone in the population. This represents the best case scenario for indirect reciprocity, making the model identical to the *n*-person direct reciprocity model of Boyd and Richerson (1988). The cost of helping, c, is set to 1, and the benefit of receiving help, b, is set to 2, meaning that an individual receives a benefit of 2 if everyone in the group, including himself, cooperates. A b/c ratio of 2 represents the minimum requirement for altruism between full siblings according to Hamilton's rule. The curves represent the required relatedness as a function of the strength of reciprocity for different groups sizes (N = 2, 4, 8, 16). When groups are dyads (N = 2), we recover Axelrod and Hamilton's (1981) model of direct reciprocity; as the shadow of future increases, the minimum relatedness required for cooperation to evolve diminishes to low levels. As group size increases, cooperation based on reciprocity will not evolve unless relatedness is quite high.

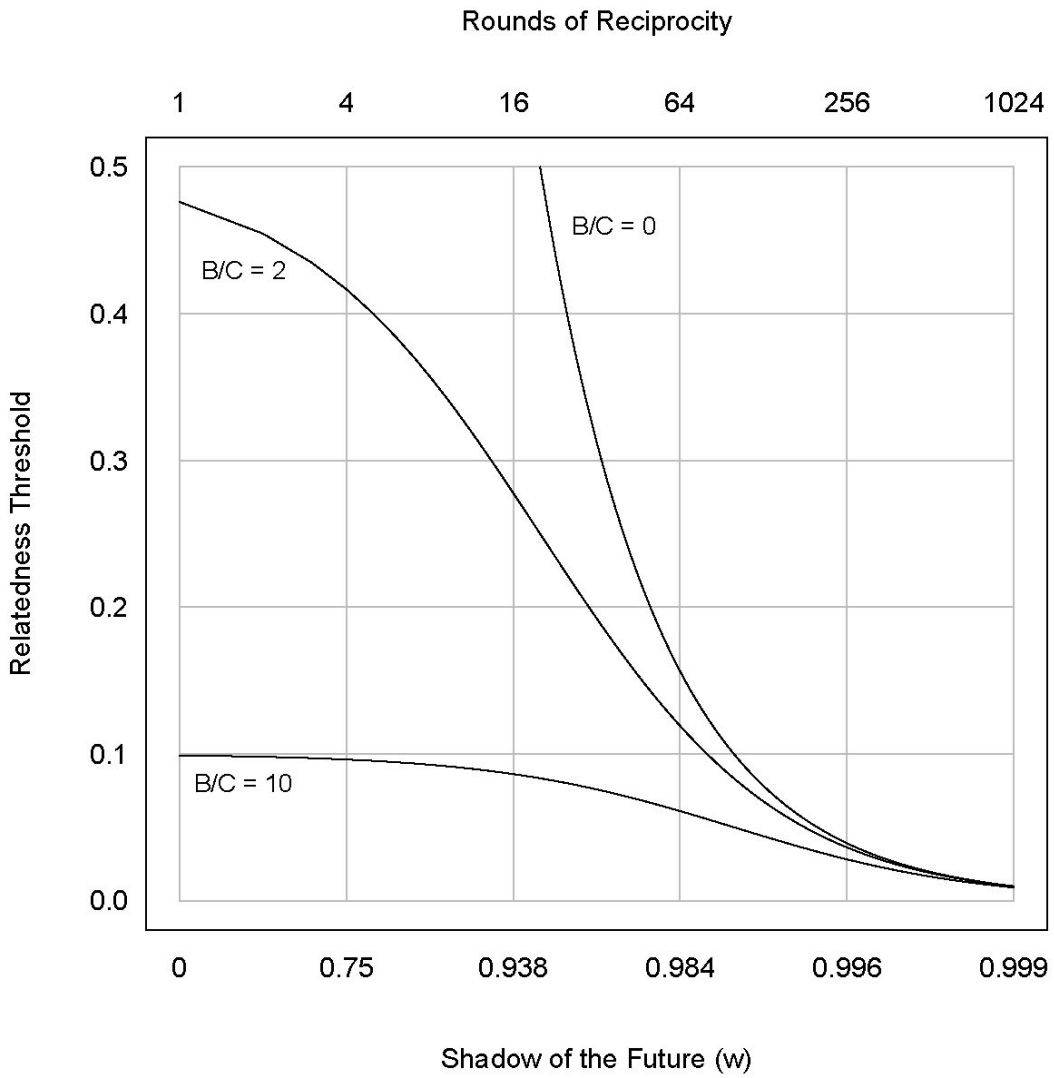

Figure 3. The minimum relatedness necessary for the *shunner* strategy to invade a population of *defectors* in a model linking dyadic indirect reciprocity and collective action. The strength of reciprocity, w, Axelrod's (1984) "shadow of the future," is depicted on the bottom x-axis on a log scale. The top x-axis shows the corresponding number of rounds of social exchange (e.g., w=0.75 implies 4 expected rounds). The reputation quality (q) is set to 1, meaning that individuals can accurately track the reputation of everyone in the population. Group sizes are large ($n >> 2$). The cost of helping during a bout of mutual aid, c, is set to 1, and the benefit of receiving help from one person during a bout of mutual aid, b, is set to 2. A b/c ratio of 2 represents the minimum requirement for altruism between full siblings according to Hamilton's rule. The cost of contributing to the collective action, C, is set to 10, implying that this cost will not be recouped until 10 rounds of mutual aid have passed. The curves represent the required relatedness as a function of the strength of reciprocity for different benefit to cost ratios for the collective action (B/C). When $B/C = 0$, the collective action is maladaptive, requiring a cost but returning no beneft. When $B/C = 2$ or 10, the collective action is welfare-enhancing, returning twice or ten times the investment. When there are few bouts of mutual aid, only highly beneficial collective actions can evolve with low levels of relatedness. As the shadow of the future increases, the relatedness threshold diminishes to low values, regardless of the effect of the collective action.

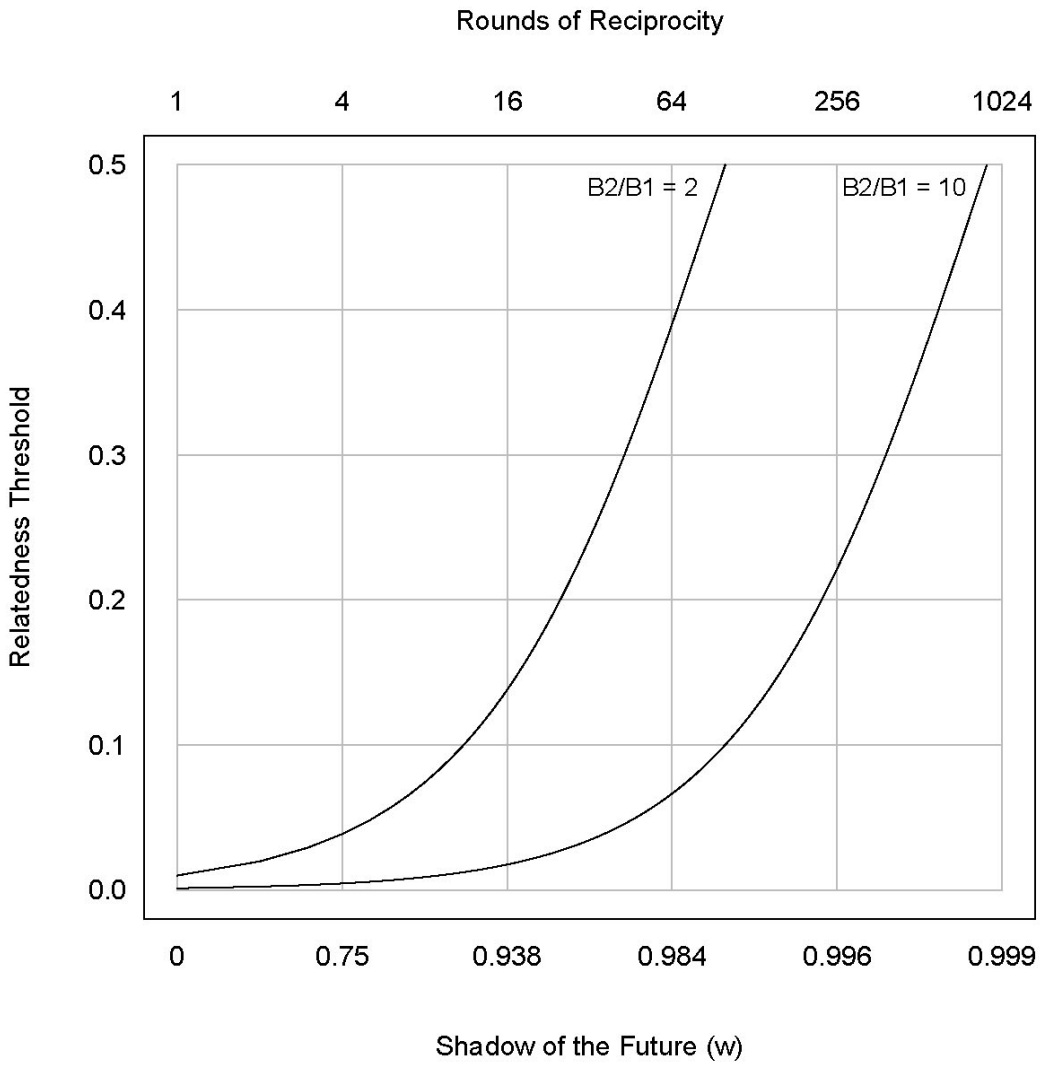

Figure 4. The minimum relatedness necessary for a *shunner* strategy practicing one collective action to invade a population of *shunners* practicing a different collective action in a model linking dyadic indirect reciprocity and collective action. The strength of reciprocity, w, Axelrod's (1984) "shadow of the future," is depicted on the bottom x-axis on a log scale. The top x-axis shows the corresponding number of rounds of social exchange (e.g., w=0.75 implies 4 expected rounds). The reputation quality (q) is set to 1, meaning that individuals can accurately track the reputation of everyone in the population. Group sizes are large ($n >> 2$). The cost of helping during a bout of mutual aid, c, is set to 1, and the benefit of receiving help from one person during a bout of mutual aid, b, is set to 2. A b/c ratio of 2 represents the minimum requirement for altruism between full siblings according to Hamilton's rule. The cost of contributing to either collective action, C, is set to 10, implying that this cost will not be recouped until 10 rounds of mutual aid have passed. The benefit of the entrenched collective action, $B1$, is set to 100, meaning that it returns 10 times the cost of contribution. The curves represent the required relatedness as a function of the strength of reciprocity for different ratios of the benefit of the new collective action, $B2$, to the benefit to existing collective action, $B1$. When $B1/B2 = 2$ or 10, the new collective action norm returns twice or ten times the benefit of the entrenched norm. When there are few bouts of mutual aid, the new collective action can invade with low relatedness. However, when there are so few bouts of mutual aid, reciprocity itself won't evolve; defection will dominate. As the shadow of the future increases, the relatedness threshold needed to supplant the current collective action increases, meaning that reciprocity and relatedness oppose one another.