# The Evolution of Strong Reciprocity[*]

Samuel Bowles
Herbert Gintis
University of Massachusetts, Amherst

July 30, 1998

### Abstract

Where genetically unrelated members of a group benefit from mutual adherence to a social norm, agents may obey the norm and punish its violators, even when this behavior cannot be justified in terms of self-regarding, outcome-oriented preferences. We call this *strong reciprocity.* We distinguish this from *weak reciprocity*, namely reciprocal altruism, tit-for-tat, exchange under complete contracting, and other forms of mutually beneficial cooperation that can be accounted for in terms of self-regarding outcome-oriented preferences We review compelling evidence for the existence and importance of strong reciprocity in human society. However, where benefits and costs are measured in fitness terms and where the relevant behaviors are governed by genetic inheritance subject to natural selection, it is generally thought that, as a form of altruism, strong reciprocity cannot invade a population of non-reciprocators, nor can it be sustained in a stable population equilibrium. We show that this is not the case, and offer an evolutionary explanation of the phenomenon.

As the late Pleistocene is the only period long enough to account for a significant development in modern human gene distributions, we base our model on the structure of interaction among members of the small hunter-gatherer bands that constituted most of the history of *Homo sapiens*, as revealed by historical and anthropological evidence.

## 1 Introduction

Where genetically unrelated members of a group benefit from mutual adherence to a social norm, agents may obey the norm and punish its violators, even when this behavior cannot be justified in terms of self-regarding,

[*] For presentation at the Santa Fe Institute, August 6-8, 1998. We would like to thank the MacArthur Foundation for financial support.

outcome-oriented preferences. We call this *strong reciprocity.* We distinguish this from *weak reciprocity*, namely reciprocal altruism, tit-for-tat, exchange under complete contracting, and other forms of mutually beneficial cooperation that can be accounted for in terms of self-regarding outcome-oriented preferences. Compelling evidence for the existence and importance of strong reciprocity comes from controlled laboratory experiments, particularly the study of public goods, common pool resource, trust, ultimatum, and other games, from the ethnographic literature on simple societies, from historical accounts of collective action, as well as from everyday observation. Strong reciprocity confers group benefits because the prospect of punishment by reciprocators reduces free riding. However reciprocity imposes individual costs, both because reciprocators contribute more to the group than non-reciprocators, and because they sustain the costs of punishing norm violation. Thus where benefits and costs are measured in fitness terms and where the relevant behaviors are governed by genetic inheritance subject to natural selection, it is generally thought that, as a form of altruism, strong reciprocity cannot invade a population of non-reciprocators, nor can it be sustained in a stable population equilibrium. We show that this is not the case, and offer an evolutionary explanation of the phenomenon.

Aspects of strong reciprocity are surely socially learned and culturally transmitted in contemporary societies. We do not seek to determine the relative importance of genes and culture in the apparent evolutionary success of strong reciprocity, but rather to answer the question: could such behavior have a genetic basis beyond the obvious requirements on the cognitive capacities of individuals.

As the late Pleistocene is the only period long enough to account for a significant development in modern human gene distributions, we base our model on the structure of interaction among members of the small hunter-gatherer bands in this period, which constitutes most of the history of *Homo sapiens*, as revealed by historical and anthropological evidence.[1]

Here we propose an explanation based on the fact that strong reciprocity supports high levels of mutual monitoring within groups, and for this reason groups with large numbers of reciprocators have superior average levels of fitness. Despite the individually costly nature of monitoring, strong reciprocity can then evolve because of the greater likelihood that reciprocators will be in groups with effective mutual monitoring. This greater likelihood derives

[1] As the mechanics of genetic determination and its associated inheritance process are not germane to our model, we leave this issue unexplored, assuming that offspring are clones of a single parent.

from the fact that norm violators, reciprocators and non-reciprocators alike, are occasionally ostracized, and non-reciprocators are more likely to be norm violators. Formally, this is an assortative interaction model characterized by a stage game supporting a stationary distribution of types within groups.

We provide a population-level equilibrium in which strong reciprocity persists even though non-reciprocators have greater fitness when interacting with reciprocators, and non-reciprocators may form a considerable fraction of the population (about 40% in a simulation we discuss below). We are also able to offer a plausible account of the successful invasion and diffusion of reciprocity behaviors a population of non-reciprocators. Under assumptions which we think may reflect the relevant historical conditions, the model thus describes the genetic evolution of reciprocal preferences.

Our model has several characteristics distinguishing it from most other accounts of reciprocity. First, we treat social interactions as completely decentralized, so we do not include third-party enforcement, for instance by a state or judicial system. Therefore the standard general equilibrium models of exchange under complete contracting (Walras 1954[1874], Arrow and Debreu 1954, Arrow and Hahn 1971) do not apply.

Second, while most decentralized models of reciprocity use repeated interactions among single pairs of agents to induce cooperative behavior (Axelrod and Hamilton 1981, Axelrod 1984, Boorman and Levitt 1980, Boyd and Lorberbaum 1987, Kreps, Milgrom, Roberts and Wilson 1982, Guttman 1996), we treat social interaction as a one-time event, or a series of one-time events in which no new knowledge is acquired from the events of the previous periods, and we assume that relatively large groups of agents interact. As Boyd and Richerson (1988) show, not even repeated interactions allows for the evolution of reciprocity in large groups of self-regarding agents.

Third, unlike Hamilton (1964), most of the sociobiological literature (Wilson 1975), and economists' contributions (Samuelson 1983, Bergstrom and Stark 1993, Bergstrom 1995), kin-selection and inclusive fitness play no role in our argument. In its place we use assortative interactions that favor reciprocators over non-reciprocators.

Fourth, unlike other models that use assortative interactions (Wilson 1980, Wilson and Sober 1994, Güth and Yaari 1992, Güth 1995, Wilson and Dugatkin 1997) we do not assume reciprocators can be distinguished from non-reciprocators by some phenotypic trait, nor can individuals establish reputations by their behaviors. Indeed, not even norm violation identifies an agent as a non-reciprocator, since reciprocators as well as non-reciprocators violate norms, and they receive the same punishments. Rather, in our model reciprocators are more likely to be in groups with other reciprocators because they have a lower frequency of norm violation, and hence are less likely to be ostracized for misbehavior.

Fifth, our treatment does not suffer the twin disability characteristic of many models of group selection of genetically transmitted traits, namely that within-group selection typically outpaces between-group selection due to the limited extent of group extinctions and the fact that within group variances of traits are typically vastly larger than between group variances (Williams 1966, Crow and Kimura 1970). These difficulties do not arise in our model because our groups exhibit a stable distribution of types across time, so the within-group selection against the group-beneficial trait does not operate.

Finally, the behaviors we seek to explain, while formally altruistic—that is individually costly and group beneficial—are more punishing than kind, reflecting Robert Trivers' (1971):49 "moralistic aggression" and Donald Campbell's (1983):37 "mutual monitoring, forcing altruism on fellow group members who cannot survive without cooperative group membership" more than the generic "mutual aid feeling" stressed by Peter Kropotkin (1989[1903]):277 and many subsequent authors.[2]

The reader knowledgeable in the biological literature may be concerned that ours is a group selection model of genetic evolution, while such models have been widely rejected by biologists as empirically implausible (Williams 1966, Eshel 1972, Maynard Smith 1976, Boorman and Levitt 1980, Dawkins 1989). In recent years, however, this assessment has been largely reversed because in the 1980's biologists developed theoretically coherent and empirically plausible models of group selection (Harpending and Rogers 1987, Uyenoyama and Feldman 1980).[3] Moreover, as William Hamilton stressed (Hamilton 1975) common ancestry is relevant to the question of altruism only insofar as it leads to assortative interactions. In particular, it is quite mistaken to consider kinship mechanisms as more basic or more 'biological' than other means of ensuring assortative interactions (Wade 1985, Breden 1990, Wilson and Dugatkin 1997, Gintis forthcoming).

[2] Our paper shares this characteristic with Sethi and Somanathan (1996). In contrast to their model, however, our reciprocators do not use weakly dominated strategies, so our model can support a positive (indeed, quite high) level of non-cooperation in equilibrium. We believe a high frequency of non-cooperation is in fact found in both simple and contemporary societies.

[3] The long-standing debate over the possibility of group-level selection effects appears to be over. For a review of the literature by strong proponents of group selection, see Sober and Wilson (1998). For a review of the book by a traditional opponent of the notion, see Maynard Smith (1998), who agrees that "selection between groups can be more effective than individual selection in producing change."

We begin in the next section by reviewing the experimental, historical, and ethnographic evidence for the importance of strong reciprocity and identify the ecological settings we wish to model. In Section 3 we model the actions of members of a single group and define a set of Nash equilibria representing their behaviors. We then turn from the *behaviors* of members within groups to their *reproductive success*, addressing in Section 4 the rate of change of genetically different types within groups and in Section 5 the distribution of types in the larger population. We then we ask if this model might explain the evolution of strong reciprocity among the hunger-gatherer foragers of the late Pleistocene.

## 2 Reciprocity in Simple and Complex Societies

The commonly observed rejection of substantial positive offers in experimental ultimatum games is our first piece of evidence for the existence of strong reciprocity. Experimental protocols differ, but the general structure of the ultimatum game is simple. Subjects are paired, one is the responder, the other the proposer. The proposer is provisionally awarded an amount ('the pie') to be divided between proposer and responder. The proposer offers a certain portion of the pie to the responder. If the responder accepts, the responder gets the proposed portion, and the proposer keeps the rest. If the responder rejects the offer both get nothing.[4] In experiments conducted in the United States, Slovakia, Japan, Israel, Slovenia, Germany, Russia, Indonesia, some with very high stakes, the vast majority of proposers offer between forty and fifty per cent of the pie, and offers lower that thirty per cent are often rejected (Fehr and Schmidt 1997). Because these behaviors occur in single-shot interactions and on the last round of multi-round interactions, they cannot be accounted for by the responder's attempt to modify subsequent behavior of the proposer. Punishment *per se* is the most likely motive.[5]

[4] See Güth, Schmittberger and Schwarz (1982), Ostrom, Walker and Gardner (1992), Güth and Ockenfels (1993), Forsythe, Horowitz, Savin and Sefton (1994), Cameron (1995), Hoffman, McCabe and Smith (April, 1996), and Falk and Fischbacher (1998). For an overview of the studies in this area, see Davis and Holt (1993) and Fehr, Gächter and Kirchsteiger (1997).

[5] As evidence for this interpretation, we note that the rejection of positive offers is substantially less when the game is altered so that rejection does not punish the proposer (Abbink, Bolton, Sadrieh and Tang 1996, Abbink et al. 1996). Moreover the fact that offers generated by a computer rather than another person are significantly less likely to be rejected suggests that those rejected offers at to cost to themselves are reacting to violations of norms rather than simply rejecting disadvantageous offers (Blount 1995).

More directly analogous to strong reciprocity in groups however, are findings in $n$-player public goods experiments. The following is a common variant. Ten players are given $1 in each of ten rounds. On each round, each player can contribute any portion of the $1 (anonymously) to a 'common pool.' The experimenter divides the amount in the common pool by two, and gives each player that much money. If all ten players are cooperative, on each round each puts $1 in the pool, the experimented divides the $10 in the pool by two, and gives each player $5. After ten rounds of this, each subject has $50. By being selfish, however, each player can do better. By keeping the $1, the player ends up with $10, plus receives $45 as his share of the pool, for a total of $55. If all behave this way, each receives $10. Thus this is an 'iterated prisoner's dilemma' in which self-regarding players all contribute nothing.

However only a small fraction of players contributing nothing to the common pool. Rather, in the early stages of the game, people generally contribute half their money to the pool. In the later stages of the game, contributions decay until at the end, they are contributing very little. However if players are permitted to retaliate directly against non-contributors, but at a cost to themselves, they do so (Fehr and Gächter 1996). In this situation, contributions rise in subsequent rounds to near the maximal level. Moreover in the experiments of Fehr and Gächter, punishment levels are undiminished in the final rounds, suggesting that disciplining norm violators is an end in itself and hence will be exhibited even when there is no prospect of modifying the subsequent behavior of the shirker or potential future shirkers.

Such experiments show that agents are willing to incur a cost to punish those whom they perceive to have treated them, or a group to which they belong badly.[6] Also in everyday life, we see people consumed with the desire for revenge against those who have harmed them or their families, even where no material gain can be expected even where such behavior cannot improve one's bargaining situation in future interactions (Nisbett and Cohen 1996, Boehm 1984).

Evidence of an entirely different type comes from the historical and contemporary study of collective action, where strong reciprocity may help to explain the occasional participation of large numbers of people in insurrectionary and protest activities that on balance are individually costly, though

[6] See Ostrom et al. (1992) on common pool resources, Fehr et al. (1997) on efficiency wages, and Fehr and Gächter (1996) on public goods. Coleman (1988) develops the parallel point that free riding in social networks can be avoided if network members provide positive rewards for cooperating.

socially beneficial. While evidence on why people participate is open to conflicting interpretations, and direct material incentives are often present, the important role for a desire to punish those who have violated norms or harmed others seems inescapable in many cases.

Barrington Moore, Jr. (1978) sought common motivational bases—"general conceptions of unfair and unjust behavior" (21)—for the moral outrage fueling struggles for justice that have recurred throughout human history. "There are grounds," he concludes from his wide-ranging historical investigation,

> for suspecting that the welter of moral codes may conceal a certain unity of original form… a general ground plan, a conception of what social relationships ought to be. It is a conception that by no means excludes hierarchy and authority, where exceptional qualities and defects can be the source of enormous admiration and awe. At the same time, it is one where services and favors, trust and affection, in the course of mutual exchanges, are ideally expected to find some rough balancing out. (4-5,509)

Moore termed his discovery "… the concept of reciprocity—or better, mutual obligation, a term that does not imply equality of burdens or obligations…"(506) In like manner James Scott (1976) analyzed agrarian revolts identifying violations of the "norm of reciprocity" as one the essential triggers of insurrectionary motivations. We do not think that Scott's or Moore's assessments are idiosyncratic.

Finally we come to the ethnographic evidence concerning strong reciprocity in simple societies, by which we mean small scale societies with little formal differentiation among adult males. The widespread sharing of food, valuable information, and other sources of survival among many of these societies is well established (Woodburn 1982). Strong reciprocity, including spontaneous sharing and the sanctioning of those who violate sharing norms, provides a parsimonious explanation. Punishing norm violators deters free riding and hence explains both sharing and but working to acquire goods that later would be shared.[7]

[7] Sharing may be unavoidable when large game is hunted and meat cannot be stored (a single antelope hunted by the !Kung may provide 300 person-days of food (Lee 1979):435). Nicholas Blurton-Jones (1987) has made this fact central to his "tolerated theft" model of sharing in simple societies according to which food will be shared to the extent that fitness benefits are concave in food and food is acquired irregularly in large packages. But food sharing is widespread even where storage is an option, for example among the Inuit and other Eskimos (Damas 1972, Wenzel 1995, Hawkes 1993), and even in the Kalahari where

If Woodburn's (1982):431 characterization of these societies as examples of "politically assertive egalitarianism" is correct, the evolutionary puzzle is not why they share or even why they produce goods destined for sharing—punishment of those who violate a sharing or working norm could readily make working and sharing a fitness-enhancing strategy—but rather why they punish. To explore this question we develop a model of norm adherence and the punishment of norm violators.

Because it is insufficient to explain sharing goods without explaining why individuals would work to acquire goods knowing they would be shared, we will focus on the team production case. To do this we develop a model in which it is costly to following a work norm and costly to punish norm violators. Could those whose strong reciprocity induce such costly behaviors evolve under the influence of natural selection?

Our model captures key of characteristics of small foraging bands.[8] First, groups are sufficiently small that members directly observe and interact with one another, yet sufficiently large that the problem of free riding in team production is present. Second, there is no centralized structure of governance (state, judicial system, Big Man, or other) so the enforcement of norms depends on the voluntary participation of peers. Third, there are many unrelated individuals, so altruism cannot be explained by inclusive fitness. Fourth, status differences are quite limited, especially by comparison to horticultural and later industrial societies, which justifies our treatment of individuals as homogeneous other than by reciprocator/non-reciprocator type and by the group to which they belong. Fifth, the sharing on which our model of team production is based—either of food individually acquired or of the common work of acquiring food—seems likely to have been characteristic of these societies. Sixth, hunter-gather bands experience high membership turnover, justifying our abstraction from reputation effects and repeated interactions as means of norm enforcement. Finally the only intertemporal relationships in our model concern fitness: the individuals in our model do not invest—store food or accumulate resources—and this too is a characteristic of at least those hunter-gather bands based on what Woodburn (1982)

meat drying is known but not widely practiced (Cashdan 1980). But more importantly, as Kristen Hawkes (1993) observed, forced sharing merely displaces the evolutionary puzzle from "why to they share" to "why do they hunt?"

[8] We have relied on the following sources: Woodburn (1982), Boehm (1982), Boehm (1993), Blurton-Jones (1987), Cashdan (1980), Knauft (1991), Hawkes (1992), Hawkes (1993), Kaplan and Hill (1985b), Kaplan and Hill (1985a), Kaplan, Hill, Hawkes and Hurtado (1984), Lee (1979), Woodburn and Barnard (1988), Endicott (1988), Balikci (1970), Kent (1989), Damas (1972), Wenzel (1995), Knauft (1989)..

calls an "immediate return" system of production.

## 3 Equilibrium Working, Shirking and Punishing Within a Group

Consider a group with $n$ members, the size of which is determined exogenously. If members work non-reciprocally, they all have equal fitness $\phi_a$, which we define as the number of replicas produced per individual minus one, or equivalently, the rate of growth the population in question. We assume $\phi_a < 0$, so a group loses members over time if its members behave non-reciprocally. If the members work reciprocally, however, if there is no shirking, and if output is divided equally among members, each will have positive fitness.

However group members benefit from 'free riding' on the group—not working, while still sharing equally in the total production of the group. To model this, we suppose each member can either supply one unit of effort (work), or supply zero units of effort (shirk). Let $\sigma_j$ be the probability that member $j$ shirks, so $\sigma = \sum_{j=1}^{n} \sigma_j / n$ is the average rate of shirking. We assume output is additive over group members, so the fitness value of group output is $n(1-\sigma)q$, where $q$ is the output of one working member. We explore the case where output is shared equally, so each member gets $(1-\sigma)q$. The loss to the group from one member shirking is $q$, while the gain to a member is the fitness cost of effort, $b > 0$, which we assume is identical for all group members, and $q > b$.

We assume that $n$ is sufficiently large that $q/n < b$, so if there is no policing of free riders, shirking would promote a member's fitness whether or not the other members work or shirk.[9] However we suppose that a group member can be monitored by other members of the group, and if detected shirking, can be punished. Suppose the cost to a member of monitoring another member is $c > 0$ and a shirking member who is monitored will be detected and punished with probability $p \in (0, 1]$. Punishment consists of sustaining a cost $s > 0$, and being ostracized from the group.[10]

We now face a 'second order free rider problem': it is costly to monitor and to punish, so each member would like the others to monitor and punish, but suffers material losses by doing so himself. Suppose, however, the group consists of two type of actors. The first type maximizes fitness, and therefore never punishes, and only works if the cost of being detected

[9] We abstract here and below from inclusive fitness considerations, so helping others cannot directly promote the fitness of one's genes.

[10] Dugatkin (1979) suggests the possibility that a group selection model might be based on exiling non-cooperators.

and punished is sufficiently high that the fitness costs of shirking exceed the fitness benefits. The second type, whom we call *reciprocators*, are motivated not only by fitness considerations, but also a subjective utility $\rho > 0$ from punishing those who violate group norms, as well as by a concern for the well-being of other reciprocators. To capture the latter, we assume reciprocators experience a disutility of labor that is declining in the fraction $f$ of the group which is reciprocators or, for simplicity, $b - f\epsilon$.[11] We assume both types are homogeneous. We assume throughout that $\rho p > c$, so the expected subjective benefits from punishing shirkers exceed the cost of monitoring a shirker. We call the fitness-maximizers *non-reciprocators*.

The introduction of reciprocators solves the second order free rider problem only by displacing it to the following question: How might the behaviors associated with preferences that are not fitness-maximizing—namely those associated with $\rho$ and $\epsilon$—have evolved under the influence of natural selection operating on genetically transmitted traits? To answer this we explore whether individuals with these preferences might enjoy an average level of fitness as great as the fitness-maximizing non-reciprocators. Thus, we will have to distinguish between individual utilities, which regulate behaviors, and levels of fitness, which determine the evolution of the composition of the population. When we refer to payoffs, we mean the utilities, which only in the case of non-reciprocators is equivalent to fitness.

We assume a non-reciprocator cannot be distinguished from a reciprocator. While the act of shirking is observable, the type of a shirker need not be deducible therefrom.[12] Moreover, since shirkers are ostracized, members do not accumulate information concerning other members's behavior in previous periods, so we are free to assume that all group members are monitored equally. Moreover, our homogeneity assumptions imply that there will be a common rate of monitoring $\mu$ in equilibrium for all reciprocators, while non-reciprocators do not monitor. There will also be a common rate of shirking $\sigma_r$ for reciprocators and $\sigma_n$ for non-reciprocators, so if the proportion of reciprocators is $f$, we have

$$\sigma = f\sigma_r + (1-f)\sigma_n. \qquad (1)$$

If a reciprocator monitors, the likelihood of detecting a non-reciprocator

[11] We assume $f$ is common knowledge, but group members cannot tell the type of individual fellow members. Our model is changed little if we assume the disutility of labor is simply $b - \epsilon$.

[12] As we shall see, there are some equilibria of our model in which shirking uniquely identifies a member as a non-reciprocator, but in the equilibria of interest, this is not the case.

shirking is $\sigma_n p$, and the corresponding likelihood for a reciprocator is $\sigma_r p$ so the expected net cost of monitoring is[13]

$$c - (p\rho f\sigma_r + p\rho(1-f)\sigma_n) = c - p\sigma\rho. \tag{2}$$

For simplicity, we assume the probability that a shirker is detected not working when each of the reciprocators monitors at rate $\mu$ is linear in total monitoring, and so equals $fn\mu p$. Writing the gain to a non-reciprocator from shirking as the cost of working minus the foregone share of output, we have

$$g_n = b - \frac{q}{n}. \tag{3}$$

Given $s$, we can write the expected non-reciprocator gain from shirking as $g_n - fn\mu ps$. Reciprocators gain $b - q/n - \epsilon f$ from shirking, so $g_r = g_n - \epsilon f$. Then if $\sigma_n$ and $\sigma_r$ are chosen by reciprocators and non-reciprocators as best responses, we have

$$\sigma_n \begin{cases} = 0, & g_n < fn\mu ps \\ \in [0,1], & g_n = fn\mu ps \\ = 1, & g_n > fn\mu ps \end{cases} \qquad \sigma_r \begin{cases} = 0, & g_r < fn\mu ps \\ \in [0,1], & g_r = fn\mu ps\;. \\ = 1, & g_r > fn\mu ps \end{cases} \tag{4}$$

We assume that $g_r < 0$ when $f = 1$, so that universal cooperation holds in a group of reciprocators with no monitoring. However for $f < 1$, any Nash equilibrium involves positive shirking, since if $\sigma = 0$ then $p\rho\sigma < c$, so $\mu = 0$, so the cost of shirking is zero, and since $g_n > 0$, it follows that $\sigma_n = 1$ so $\sigma > 0$ by (1), which is a contradiction. Thus we must investigate conditions under which $0 < \sigma < 1$ in equilibrium. We have

**Theorem 1.** *Suppose $p\rho > c$, $b > q/n$, and $b - q/n < \epsilon$. The following cases are nonempty and exhaustive.*[14]

*(a) If $f < g_n/(nps + \epsilon)$ then $\sigma = \sigma_r = \sigma_n = 1$ and $\mu = 1$. This is a 'sado-masochistic' equilibrium in which all members shirk, and nevertheless which reciprocators monitor with certainty.*

[13] To be exact, we should take into account the fact that no member can monitor himself, so the correct formula is

$$c - p\left(\overline{f}\sigma_r + (1-\overline{f})s_n\right)\rho.$$

where $\overline{f} = (fn-1)/n$. To simplify the exposition, however, we will assume $n$ is sufficiently large that we can replace $\overline{f}$ by $f$ in our calculation, after which the previous expression simplifies to the above expression.

[14] As mentioned in the text, the first inequality implies that reciprocators monitor when the probability of detecting shirking is unity; the second says that the fitness cost of working is greater than the fitness payoff to working; the third says that in a group of all reciprocators without monitoring, there is zero shirking.

*(b) If $f < g_n/nps$, $f < 1-c/p\rho$, and either $g_n/(nps+\epsilon) < f$ or $g_n/\epsilon < f$, then $\sigma_n = 1$, $\sigma_r = 0$, $\sigma = 1-f$ and $\mu = 1$. This is an equilibrium in which non-reciprocators surely shirk, reciprocators never shirk, and reciprocators monitor with certainty.*

*(c) If $g_n/nps < f < 1-c/p\rho$, then $\mu = g_n/fnps$, $\sigma_r = 0$, $\sigma_n = c/p\rho(1-f)$, $\sigma = c/p\rho$. In this equilibrium reciprocators still never shirk, but non-reciprocators work, and reciprocators monitor, with positive probability.*

*(d) If $g_n/(nps+\epsilon) < f < g_n/\epsilon$ or $g_n/\epsilon < f < g_n/\epsilon$, and $f > 1-c/p\rho$, then $\sigma_n = 1$,*

$$\sigma_r = \frac{1}{f}\left[\frac{c}{p\rho} - (1-f)\right], \tag{5}$$

*so $\sigma = c/p\rho$ and $\mu = g_r/fnps$. Here non-reciprocators do not work and reciprocators have positive shirking, while monitoring with less than certainty.*

*(e) If $g_n/\epsilon < f$, $f > 1-c/p\rho$ then $\sigma_n = 1$, $\sigma_r = 0$, $\sigma = 1-f$ and $\mu = 0$. In this equilibrium reciprocators never shirk and never monitor, while non-reciprocators shirk with certainty.*

The proof is in Appendix A. The intuition underlying the proof is illustrated by the depiction of the theorem's five cases in Figure 1. For case (a), where values of $f$ are less than $g_n/(nps+\epsilon)$, the payoff to shirking for the reciprocators exceeds the expected cost when all reciprocators monitor (4), so reciprocators shirk and, a fortiori, so do non-reciprocators. For $f$ slightly larger than this value (if the cost of monitoring is low), we have case (b), where all reciprocators work and continue to monitor while for $f > g_n/nps$ we have case (c), where non-reciprocators also work, while by (27) the overall reduction in shirking induces reciprocators to reduce their level of monitoring. However if the cost of monitoring, $c$, exceeds $(1-f)/p\rho$, monitoring at level $\mu = 1$ is no longer a best response, even when, as in case (d), all non-reciprocators are shirking, so reciprocators pursue a mixed strategy with respect to both shirking and monitoring. Finally in case (e), where $f > g_n/\epsilon$, shirking is no longer a best response for reciprocators, while the remaining shirkers $(1-f)n$ are too few to motivate monitoring, so reciprocators work and do not monitor and non-reciprocators shirk.

Figure 2 illustrates the relationship between the fraction of reciprocators and the average level of shirking when monitoring costs are low ($c/p\rho < 1-g_n/\epsilon$). Notice that shirking is complete in region (a), but when $f$ moves into region (b), shirking falls discontinuously and declines monotonically

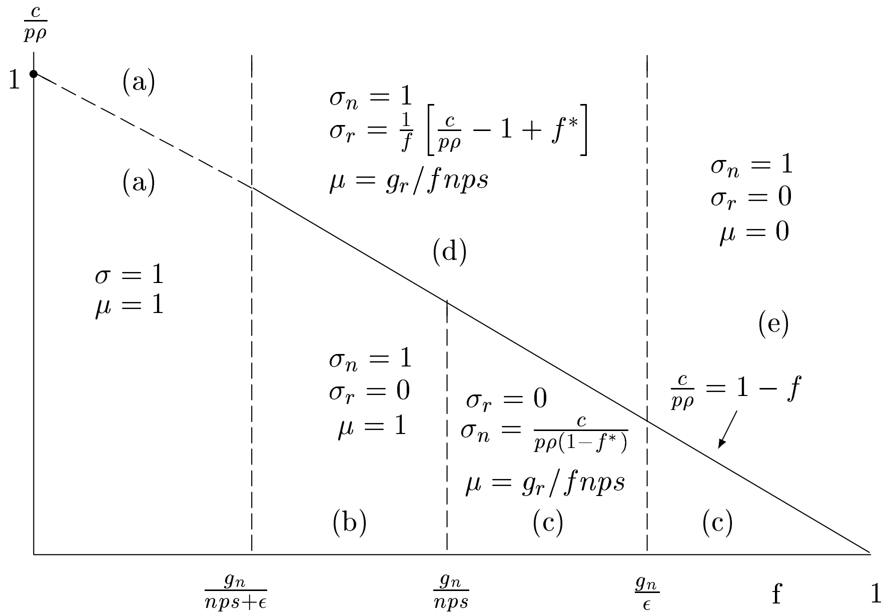

**Figure 1:** Case of Within-Group Interaction for Different Costs of Monitoring and Frequency of Reciprocators (cases (a) through (e) refer to the corresponding parts of Theorem 1). The figure assume $nps > \epsilon$, so one part of region (b) is not illustrated.

until $f$ is in region (c), after which it remains constant until it reaches region (e), where shirking falls linearly to zero as $f$ goes to 1. Figure 3 illustrates the same relationship when monitoring costs are higher ($1 - g_n/\epsilon < c/p\rho < 1 - g_n/nps$). Again shirking falls discontinuously from region to region with increasing numbers of reciprocators. The remaining cases are similar, except as $c/p\rho$ increases, first region (c) disappears, and region (b) disappears.

Figure 4 illustrates the relationship between the fraction of reciprocators on the one hand, and the monitoring rate by reciprocators and total monitoring on the other, in the case of low monitoring costs $c/p\rho$. Notice that the total material resources devoted to monitoring ($fn\mu c$) increases through regions (a) and (b), then declines as $f$ increases through the remaining regions. Similar results hold for higher values of $c/p\rho$.

## 4 Group Level Equilibrium

We have identified Nash equilibria for the behaviors of members in groups with given frequencies of reciprocators. Under what conditions will the within-group frequency of reciprocators be stationary? To explore the pop-

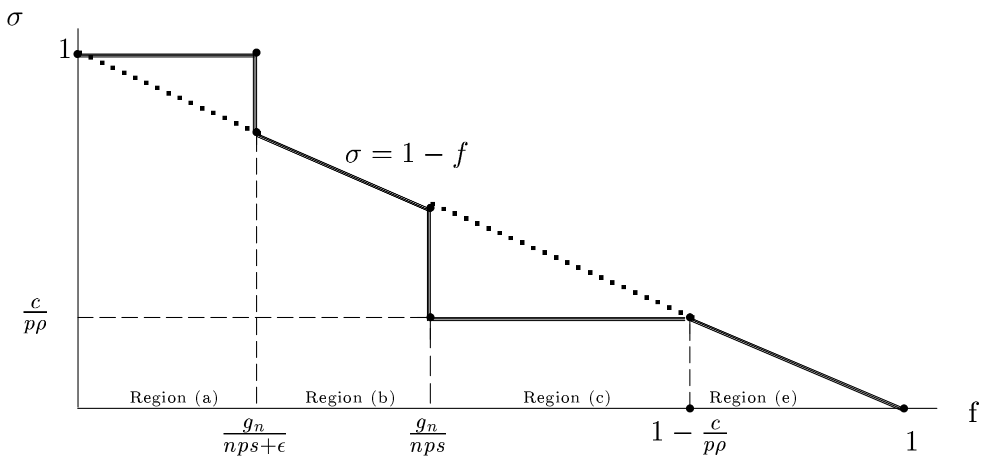

**Figure 2:** Relationship of Average Shirking to Fraction of Reciprocators with low monitoring costs ($c/p\rho < 1 - g_n/\epsilon$).

ulation dynamics within the group for the five cases identified by Theorem 1, we turn from the behavioral analysis involving utilities, to a reproduction analysis involving fitness. We will see that the frequency of reciprocators in the group may be stationary despite the greater fitness of the non-reciprocators. The reason is that while reciprocators produce more replicas, some are expelled from the group, and there is some frequency of reciprocators for which the level of ostracism is sufficient to offset the greater fitness of non-reciprocators, leading to stationarity of $f$. We will account for those ostracized subsequently, when we study to evolution of the distribution of types not in a single group, but in the population as a whole.

Let $\pi_r$ and $\pi_n$ be the rate of growth of the reciprocators and non-reciprocators in the group per time period taking account of the numbers lost through ostracism. Then if the fraction of reciprocators is $f_t$ at time $t$, at time $t + \Delta t$ is

$$f_{t+\Delta t} = \frac{f_t(1 + \pi_r \Delta t)}{f_t(1 + \pi_r \Delta t) + (1 - f_t)(1 + \pi_n \Delta t)}.$$

Subtracting $f_t$, dividing by $\Delta t$, and passing to the limit, we get

$$\frac{df_t}{dt} = f_t(\pi_r - \pi) = f_t(1 - f_t)(\pi_r - \pi_n). \tag{6}$$

Also, stability requires

$$\frac{d\pi_r}{df_t} < \frac{d\pi_n}{df_t}. \tag{7}$$

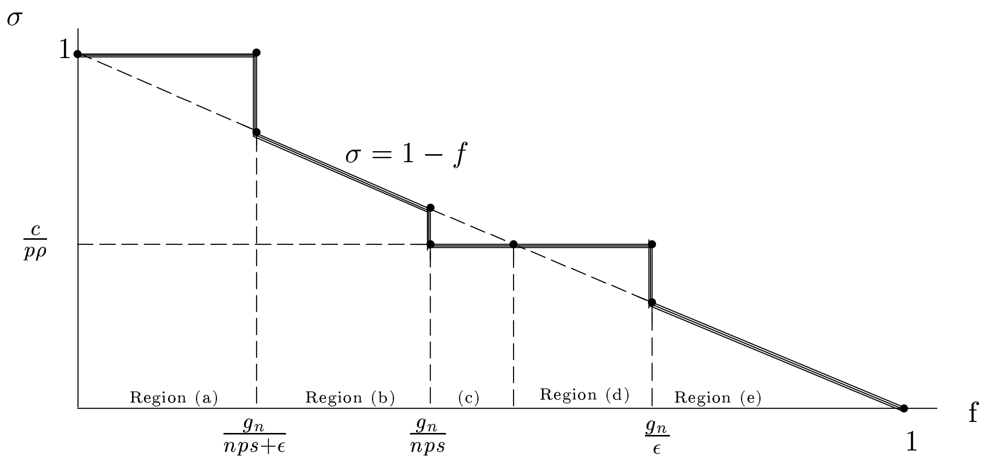

**Figure 3:** Relationship of Average Shirking to Fraction of Reciprocators with higher monitoring costs ($1 - g_n/\epsilon < c/p\rho < 1 - g_n/nps$).

In region (a) we have $\pi_n = -fnps$ and $\pi_r = -fnps - nc$. This is a bizarre 'sado-masochistic' equilibrium in which reciprocators punish even though it has no effect on the level of shirking. We henceforth assume that $-g_n < \phi_a$, so $-fnps < \phi_a$ for all $f$ in this region. Therefore both reciprocators and non-reciprocators prefer to work alone rather than in the team, with fitness payoff $\phi_a$. We take this to be the equilibrium for $f$ in this region.

In region (b), all members receive fitness benefits $qf$, reciprocators pay $b + nc$, and non-reciprocators, all of whom shirk, pay $fnps$ in punishment, plus they are ostracized at rate $fnp$. thus we have $\pi_n = qf - fnp(s+1)$ and $\pi_r = qf - b - nc$. Equating these to achieve $\dot{f} = 0$, the equilibrium condition (6) becomes

$$f^* = \frac{b + nc}{np(s+1)}.$$

The meaning of this is clear: $b + nc$ is the fitness cost of reciprocators, and $f^*np(s+1)$ is the fitness cost of non-reciprocators plus the expectation of an ostracism: equating the two gives a within-group population equilibrium. However it is readily seen that

$$\frac{d(\pi_r - \pi_n)}{df} > 0$$

for all $f$ in region (b), so $f^*$ is unstable.

In region (c) all members receive $q(1-\sigma)$, reciprocators pay $b$ in fitness costs for working and $nc\mu$ for monitoring, while non-reciprocators pay $b(1-$

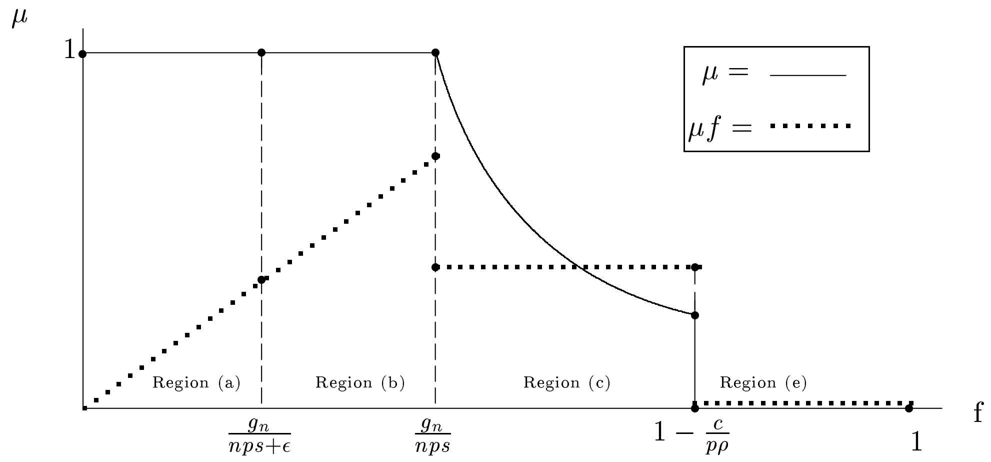

**Figure 4:** Monitoring Rate ($\mu$) and total monitoring ($f\mu$).

$\sigma_n$) for working and $fnp\mu(s+1)\sigma_n$ for shirking. Thus $\pi_r = q(1-\sigma)-b-nc\mu$ and $\pi_n = q(1-\sigma)-b(1-\sigma_n)-fnp(s+1)\sigma_n\mu$. Evaluating $d(\pi_r-\pi_n)/df$ for region (c) we find that $d\pi_r/df$ is positive: as $f$ rises, reciprocators suffer fewer monitoring costs because fewer of them monitor. After some manipulation, $d\pi_n/df$ is seen to have the sign of $[(1+s)q/n-b]$ which is negative for all but unsustainably small group sizes. To see this, note that $b > q/n$ by the assumption that shirking is a best response in the absence of monitoring, and for plausible group size $b$ will exceed $q/n$ by a considerable amount, while $s$ is not likely to exceed $1/n$, as the fitness costs imposed on a shirking member are not likely to exceed the elimination of one of the shirker's offspring. So $d(\pi_r-\pi_n)/df > 0$, and therefore, should an interior equilibrium exist in $c$, it will be unstable under reasonable assumptions. We assume that no groups exist in region $c$.[15]

In region (d), all agents receive $q(1-\sigma)$, non-reciprocators are punished at the rate $fnp(1+s)\mu = g_r(1+s)/s$, reciprocators pay $b(1-\sigma_r)$ for working, $\sigma_r g_r(1+s)/s$ for shirking, and $\mu cn$ for monitoring. Thus

$$\pi_n = q(1-\sigma)-g_r(1+s)/s, \qquad \pi_r = q(1-\sigma)-b(1-\sigma_r)-\sigma_r g_r(1+s)/s-cn\mu. \tag{8}$$

[15] There may also be a stable equilibrium at $f = g_n/nps$, the boundary between case (b) and case (c), as exhibited in Figure 6. Since $\pi_r(f)-\pi_n(f)$ is discontinuous at this point, we cannot a Jacobian argument to show this. However, clearly $f$ increases a little to the right of this boundary, and decreases a little to the left. We do not consider this to be a plausible social equilibrium, since it implies violent swings in behavior for small deviations in group composition.

We find that

$$f^* = \frac{g_n}{\epsilon} - \frac{bs(1-\sigma)}{\epsilon((1-\sigma)(1+s)-\sigma\rho)}. \tag{9}$$

Also

$$\frac{d}{df}(\pi_r - \pi_n)|_{f^*} = -\frac{\epsilon((1-\sigma)(1+s)-\sigma\rho)}{sf^*}. \tag{10}$$

Since $f^* < g_n/\epsilon$, the denominator in the second term in (9) must be positive. Therefore the numerator in 10 is positive, so the equilibrium is stable.

| Variable | Value | Description |
|---|---|---|
| $\mu$ | 0.117 | Monitoring Rate by Reciprocators |
| $\sigma_r$ | 0.344 | Shirking Rate by Reciprocators |
| $\sigma_n$ | 1.000 | Shirking Rate by Non-Reciprocators |
| $\sigma$ | 0.600 | Average Shirking Rate |
| $f^*$ | 0.610 | Frequency of Reciprocators |
| $\phi_r^s$ | 0.037 | Fitness of Reciprocators |
| $\phi_n^s$ | 0.107 | Fitness of Non-Reciprocators |
| $fnp\mu\sigma_r$ | 0.039 | Rate of Ostracism of Reciprocators |
| $fnp\mu$ | 0.107 | Rate of Ostracism of Non-Reciprocators |
| $\pi_r$ | -0.054 | Rate of growth of Non-Reciprocators in Group |
| $\pi_r$ | -0.054 | Rate of growth of Reciprocators in Group |

**Figure 5:** Equilibrium Working, Shirking, and Punishing in Region (d)
Note: The within-group equilibrium for this case is generated using the following parameter values: $n = 50$, $q = 0.15$, $s = 0.65$, $b = 0.225q$, $p = 0.03$, $\rho = 0.5$, $\epsilon = 0.039$, $c = 0.009$.

In region (e), all members receive $q(1-f)$, but reciprocators pay $b$ for working while non-reciprocators do not. Therefore $\pi_r - \pi_n = -b$ in this region, so the fraction of reciprocators declines through the region.

Figure 5 illustrates the stable equilibrium for case (d) for a particular choice of the model's parameters $(c, \rho, s, b, q, n, p, \epsilon)$. The values are stationary in two senses. First, the behaviors of the individuals are best responses and so the outcome is a Nash equilibrium for the within-group population frequency $f^*$. Second, the frequency of each type is stationary under the replicator dynamic (10), the differential fitness of the non-reciprocators being offset by their greater likelihood of being ostracized.

Figure 6 summarizes the within-group population dynamics for $f \in [0, 1]$ using the parameter values assumed in Figure 5. The equilibrium described in Figure 6 is indicated by $f^*$ in Figure 5. Using Theorem 1 we are able to prove

**Corollary 1.1.** *For plausible levels of the model's parameters ($n$, $q$, $s$, $b$,$p$, $\rho$, $\epsilon$ and $c$), a stable interior equilibrium frequency of reciprocators exists.*

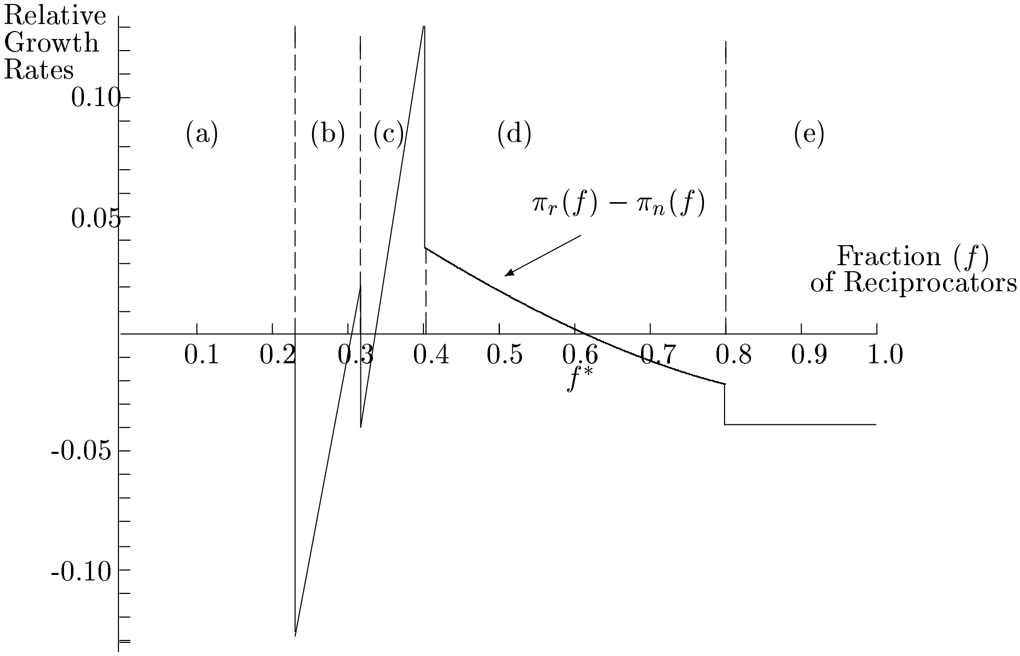

**Figure 6:** Within-Group Dynamics:$df/dt = \pi_r(f) - \pi_n(f)$, so $f = 0.61$ is a stable equilibrium, while $f = 0.30$ and $f = 0.34$ are unstable.

## 5 Population Dynamics and Equilibrium

It remains to show that $\tau$, the fraction of reciprocators in the population, has a time-invariant equilibrium value, despite the fact reciprocators have lower fitness than non-reciprocators, when the two types interact in a social group.

We suppose the total population is constant at size $N$ and that the only two types of groups are (d), which we call *social* and (a), which we call *asocial*. We assume social groups are in internal equilibrium as described in the previous section, with a fraction $f_s = f^*$ as defined in (9), and the asocial groups have a fraction $f_a < g_n/(nps + \epsilon)$ (see Theorem 1a) of reciprocators yet to be determined. Let $-\beta$ be the post-ostracism rate of growth of social groups, so $\beta$ is the immigration rate into social groups, as expressed in (25). We assume members ostracized from a social group migrate to asocial groups. Also, if $\beta < 0$, $|\beta| n$ members of each social group migrate to form new social groups, and if $\beta > 0$, members of asocial groups migrate back to social groups (since social groups cannot discriminate by type, we assume immigrants have the same fraction of reciprocators as the asocial groups from which they came). For simplicity of exposition, in developing the dynamical equations governing population movements we assume $\beta \geq 0$, leaving the (easier) case $\beta < 0$ aside.

We first develop a differential equation expressing the movement of $f_{a,t}$, the fraction of reciprocators in asocial groups at time $t$ (we assume all have the same composition of reciprocators and non-reciprocators), Let $\nu$ be the rate at which shirkers are ostracized from social groups, and let and $\sigma_r$ is the rate at which reciprocators shirk in social groups (all non-reciprocators shirk with certainty). Then if $\mu$ is the monitoring rate in social groups, using (9) and Theorem 1d, we have

$$\nu = f_s n p \mu = \frac{g_r}{s} = \frac{b(1-\sigma)}{(1+s)(1-\sigma) - \rho\sigma}$$

where $\sigma = c/p\rho$, and the total number of ostracized from a single group, including shirking reciprocators, is

$$n\nu[1 - f + f\sigma_r] = n\nu\sigma.$$

Let $\alpha$ be the fraction of the population in social groups. From the above, we see that at time $t + \Delta t$ the number of reciprocators in asocial groups after immigration and emigration is

$$f_{a,t}(1-\alpha)N(1+\phi_a\Delta t) + \alpha N(f_s\nu\sigma_r - \beta f_{a,t})\Delta t.$$

and the total number of members of asocial groups at time $t + \Delta t$ is given by the fitness $\phi_a$ of individual in asocial groups plus migrants ostracized from social groups minus emigrants, or

$$(1-\alpha)N(1+\phi_a\Delta t) + \alpha N(\nu\sigma - \beta)\Delta t.$$

Dividing these two expressions, subtracting $f_{a,t}$, dividing by $\Delta t$ and passing to the limit, we find that the fraction $f_a$ of reciprocators in asocial groups satisfies the differential equation

$$\dot{f}_{a,t} = -\frac{\alpha\nu\sigma}{1-\alpha}(f_{a,t} - f_a^*), \tag{11}$$

where

$$f_a^* = f_s\frac{\sigma_r}{\sigma} \tag{12}$$

is the equilibrium fraction of reciprocators in asocial groups, which requires that the ratio of reciprocators ostracized from social groups to all of those ostracized be equal to the ratio of reciprocators in asocial groups. As $\sigma_r < \sigma$, (12) shows that the equilibrium frequency of reciprocators in asocial groups is less than their frequency in social groups. The equation has the solution

$$f_{a,t} = f_a^* + (f_{a,0} - f_a^*)e^{-\frac{\alpha\nu\sigma}{1-\alpha}t}.$$

Now let $\phi_r$ and $\phi_n$ be the average fitness of reciprocators and non-reciprocators in the population. We can then assume a replicator dynamic, as in (6), now defined at the population rather than the group level. With $\tau$ representing the fraction of reciprocators in the population,

$$\frac{d\tau}{dt} = \tau(1-\tau)(\phi_r - \phi_n), \tag{13}$$

from which we see that stationarity of $\tau \in (0,1)$ requires that $\phi_r = \phi_n$; i.e., population-average fitnesses of reciprocators and non-reciprocators must be equal.

We obtain the expression for $\phi_r$ as follows. Let $\alpha_r$ be the fraction of reciprocators who are in social groups, let $\phi_r^s$ be the fitness of reciprocators in social groups. Since $\phi_a$ is the fitness of reciprocators in asocial groups, we have

$$\phi_r = \alpha_r\phi_r^s + (1-\alpha_r)\phi_a. \tag{14}$$

Similarly, if $\alpha_n$ is the fraction of non-reciprocators who are in social groups, and $\phi_n^s$ is the fitness of non-reciprocators in social groups, we have

$$\phi_n = \alpha_n\phi_n^s + (1-\alpha_n)\phi_a. \tag{15}$$

Moreover we have

$$\alpha_r = \frac{\alpha f_s}{\alpha f_s + (1-\alpha)f_a}, \qquad \alpha_n = \frac{\alpha(1-f_s)}{\alpha(1-f_s) + (1-\alpha)(1-f_a)}, \tag{16}$$

where, as before, $\alpha$ is the fraction of the total population in social groups. We also have

$$\phi_r^s = \sigma_r\nu - \beta, \qquad \phi_n^s = \nu - \beta. \tag{17}$$

Substituting these expressions in (14) and (15) and solving for the equilibrium condition $\phi_r = \phi_n$, we find that

$$\frac{\nu\sigma_n - \beta - \phi_a}{\nu - \beta - \phi_a} = \frac{\alpha_n}{\alpha_r}, \tag{18}$$

which requires that the relative fitness disadvantage of the reciprocators in social groups (expressed as the difference between the fitness of the type in social and asocial groups) be offset by the fitness disadvantage imposed on non reciprocators by their disproportionate location in asocial groups (given by (12).)

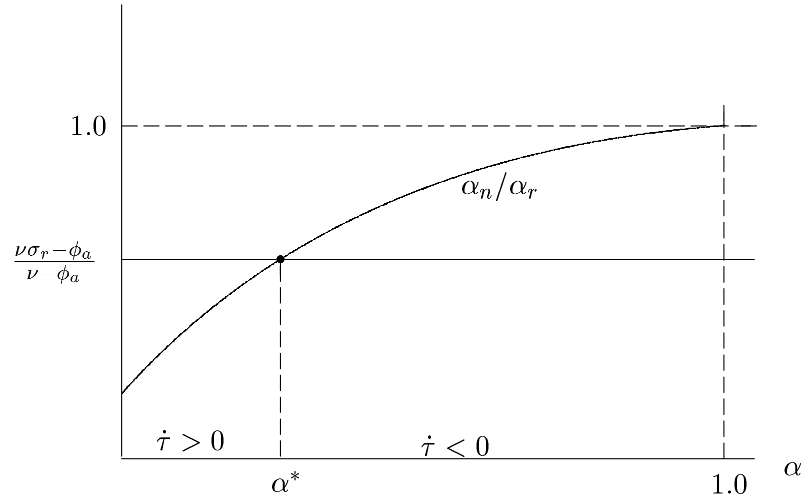

**Figure 7:** Population Level Equilibrium

The variables in (18) are completely determined by the parameter values underlying the within-group equilibrium, except for $\alpha_n/\alpha_r$, which rises monotonically in $\alpha$, attaining a value of unity when all individuals are in social groups. Thus (18) determines the equilibrium fraction of the population in social groups, as is illustrated in Figure 7. In the figure the distance $1-\alpha_n/\alpha_r$ is the advantage enjoyed by reciprocators by dint of their favorable distribution among groups, while $1-(\nu\sigma_r-\beta-\phi_a)/(\nu-\beta-\phi_a)$ is the fitness disadvantage of reciprocators in social groups arising from their propensity to engage in costly monitoring and to work hard. The equilibrium value of $\alpha$ is

$$\alpha^* = \frac{\beta+\phi_a}{\beta-\nu\sigma+\phi_a}. \tag{19}$$

Given $\alpha^*$, the equilibrium distribution of types in the population is also determined, as the distribution of types within the asocial and social groups is unchanging:

$$\tau^* = \alpha^* f_s + (1-\alpha^*)f_a.$$

Figure 8 gives the distribution of the population among groups in the population level equilibrium.

| Variable | Value | Description |
|---|---|---|
| $\tau$ | 0.540 | Population Frequency of Reciprocators |
| $\alpha$ | 0.610 | Fraction of Population in Social Groups |
| $\alpha_r$ | 0.731 | Fraction of Reciprocators in Social Groups |
| $\alpha_n$ | 0.483 | Fraction of Non-Reciprocators in Social Groups |
| $\phi_a$ | -0.100 | Fitness in Asocial Groups |
| $\phi_r^s$ | 0.037 | Fitness of Reciprocators in Social Groups |
| $\phi_n^s$ | 0.107 | Fitness of Non-Reciprocators in Social Groups |

**Figure 8:** Reciprocators and Non-Reciprocators in Population Level Equilibrium
Note: the parameter values used to generate the equilibrium are identical to those in the notes of Figure 6.

We therefore have

**Theorem 2.** *Strong reciprocators can invade a population of non-reciprocators for values such that $\alpha_n/\alpha_r > (\nu\sigma_r - \phi_a)/(\nu - \phi_a)$ evaluated at arbitrarily small $\alpha > 0$.*

We find that the stability condition for (13), which is similar to (10) is given by

$$\frac{(\beta + \phi_a)}{(\beta - \nu + \phi_a)(\nu(1-\sigma) + f_s(\beta - \nu + \phi_a))} < 0.$$

The equilibrium is thus stable if and only if

$$\nu\sigma_r - \beta > \phi_a > 0,$$

which is true as long as the fitness of social group members exceeds that of asocial group members.

We would expect that in equilibrium, the growth rate of social groups would be zero, so $\beta = 0$. To see that this is the case, we treat $\phi_r^s$ and $\phi_n^s$ as parameters, and solve (14), (15), (16) and (17) for the equilibrium condition $\dot{\phi}_r = \dot{\phi}_n$, getting

$$\alpha = \frac{f_s(\phi_r^s - \phi_a) + f_a(\phi_a - (f_s\phi_r^s + (1-f_s)\phi_n^s))}{(f_s - f_a)((f_s\phi_r^s + (1-f_s)\phi_n^s) - \phi_a)}. \tag{20}$$

The equation for zero total population growth is

$$\alpha(f_s\phi_r^s + (1-f_s)\phi_n^s) + (1-\alpha)\phi_a = 0, \tag{21}$$

which has solution

$$\alpha = \frac{-\phi_a}{f_s\phi_r^s + (1-f_s)\phi_n^s - \phi_a}. \tag{22}$$

The condition for both (21) and (22) to hold is

$$f_a = \frac{f_s\phi_r^s}{f_s\phi_r^s + (1-f_s)\phi_n^s}. \tag{23}$$

In equilibrium we have $f_a = f_a^*$ from (12) which, when substituted in (23) and simplified, give

$$\phi_r^s = \phi_n^s\sigma_r. \tag{24}$$

comparing this with (17), we see that $\beta = 0$.

The dynamics of the population frequency of reciprocators is illustrated by Figure 7. For $\alpha > \alpha^*$ the fitness disadvantage imposed on non-reciprocators by their disproportionate location in asocial groups falls short of their fitness advantages in social groups, so $d\tau/dt < 0$. When $\alpha = 1$ they suffer no fitness disadvantage due to their distribution among groups (all are in social groups), so their fitness advantage in social groups is the only selective force at work. For analogous reasons, when $\alpha < \alpha^*$ the reverse is true.

## 6 The Evolution of Reciprocity

Can this model illuminate a process by which strong reciprocity might have become common in human populations? To do so it must accomplish two things. First the model must offer a plausible account of the human interactions affecting relative fitness in the social and physical environments which were prevalent over the period during which this evolutionary process would have taken place. Second, it must show how the frequency of reciprocators could have risen from insignificance to a substantial fraction of the population under such conditions.

Do the interactions modeled here capture the relevant aspects of the social and physical environments of *Homo sapiens sapiens* during the past 200,000 to 50,000 years?[16] To answer this question we turn to recent and contemporary accounts of societies generally thought to resemble the foraging bands that were common during this period, among them the !Kung of

[16] This is the time span of anatomically modern humans reported by Klein (1989):344. Foley's (1987):22 estimate is 100,000 years. The horticultural societies that eventually replaced foraging bands almost everywhere appeared 12-10,000 years ago. Even Klein's lower limit for the appearance of modern humans leaves ample time for significant change in gene distributions to have taken place under the kinds of selection pressures at work.

Botswana and Namibia, the Ache of Paraguay, Batek of Malaysia, Hadza of Tanzania, Pandaram and Paliyan of South India, the Inuit of the Northwest territories, and the Mbuti Pygmies of Zaire. On the basis of this reading, as we have noted in Section 2, we believe that our model may be illuminating.[17]

In one respect, however, our model appears to diverge significantly from what is known about modern hunter gatherers, namely their lack of clear group boundaries and hence the extreme fluidity of group membership, as well as frequent group fissioning, sometimes as the result of within-group conflicts. However the limited migration among groups in our model (it is limited to the ostracized moving to asocial groups) is an artifact of our assumption that population growth is zero.

To see this assume, first, that except for the ostracism of shirkers from social to asocial groups, movement is entirely among similar groups, perhaps because while types cannot be recognized, one's group membership can, and social groups admit only members from other social groups. In this case the influx of migrants will be distributed by type just as the receiving social groups themselves are, so the migration process will have no effect on the equilibrium frequency within social groups, as long as it is not costly *per se* and does not, through the chance arrival of a sufficiently large number of non-reciprocator migrants from other social groups, temporarily reduce $f$ sufficiently that it is no longer in the basin of attraction of $f_s$.

If migrants may arrive also from asocial groups the analysis is more complicated but not substantively different. In this case the equilibrium distribution of those ostracized and those arriving as migrants from the asocial groups will be identical, otherwise the distribution of types in the asocial groups could not be stationary, by an extension of equation (12). As long as the numbers ostracized exceed the immigrants from the asocial groups, the workings of the model are unaffected, though the equilibrium fraction of reciprocators in the social groups will be reduced, thus attenuating, but not eliminating the group selection pressures in favor of reciprocators. The condition guaranteeing this must hold in equilibrium as long as the average fitness of members of social groups exceeds that of those in asocial groups.

To see how migration might be modeled, note that we have determined the equilibrium level $f^*$ of reciprocators in region (d) by assuming the group to be a closed system. Moreover, we determined a growth rate of the group

[17] Our main sources are listed in footnote 8. The difficulty in making inferences about simple societies during the late Pleistocene on the basis of contemporary simple societies is stressed by Foley (1987):75-78, but we do not find persuasive the view that the widespread deliberate egalitarianism of many contemporary simple societies is a recent adaptation to the encroachment of other societies on their livelihoods and territory.

given by (8). We have left its value to be determined by the baseline fitness of the group or other exogenous determinants. However, if $\pi = \pi_n = \pi_r > 0$ in equilibrium, we must have emigration, and if $\pi < 0$ we must have immigration, to maintain the group at its fixed size $n$. If there is emigration, we assume the surplus population, $\pi n$ in number combine perhaps with other such emigrants to form new social groups. But if $\pi < 0$, we must introduce a process of immigration to restore the group size.

Suppose $\pi < 0$, and let $\beta$ be the immigration rate that restores the group to size $n$, so in equilibrium we have

$$\beta = -\pi_r(f) = -\pi_n(f) > 0. \tag{25}$$

This of course will influence the equilibrium fraction of reciprocators, according to the fraction of reciprocators among immigrants. Suppose the fraction of reciprocators among immigrants is $f_a$. Then the expression $\pi_r - \pi_n$ in (6) must be replaced by

$$(\pi_r(f) + \beta f_a) - (\pi_n(f) + \beta(1 - f_a)) = \pi_r(f) - \partial_n(f) - \beta(1 - 2f_a).$$

Equation (9), noting that $f^* = f_s$ now becomes

$$f_s = \frac{g_n(1 - (1 + \rho)) - qs(1 - \sigma)/n}{\epsilon(1 + s) + s\beta(1 - 2f_a) - \sigma\epsilon\rho}. \tag{26}$$

Notice that when $\beta = 0$ this reduces to (9), in which case the denominator, and hence the numerator, are positive. Therefore if $f_a < 1/2$, an increase in $\beta$ lowers $f_s$, and conversely. This comparative static relationship can be inferred from Figure 6, noting that when $f_a < 1/2$, an increase in $\beta$ lowers $df/dt$ and thus shifts the whole curve downward, thus lowering $f_s$.

Notice also that if $\pi$ is negative and sufficiently large, and if $f_a$ is very small, no equilibrium in region (d) will exist. As long as $f_a \leq f^*$ the converse is not the case: being very productive cannot destroy the equilibrium in this region.

We now have $f_s$ as a function of $f_a$, and (12) gives us an expression for $f_a$ in terms of $f_s$. Finally, we can add the equilibrium equation

$$\pi_n(f) + \beta(1 - f_a) = 0,$$

which ensures a stable size social group. These three equations jointly determine $f_s$ and $f_a$, the fraction of reciprocators in the social and asocial groups, and $\beta$, the rate of migration into social groups.

We have not modeled group dissolution and group formation, but it is a plausible mechanism by which the distribution of the population between social and asocial groups, $\alpha$, might occur. Suppose that group size $n$ is determined in such a way that smaller or larger groups are not viable, larger ones extruding migrants and smaller ones dissolving with some probability. If $\alpha > \alpha^*$, so a greater than equilibrium number of the population are in social groups, the growth rate of the population in groups will be negative, and as groups become smaller they will be prone to dissolution, their former members dispersing to other social groups (restoring their size) or throughout the population. If, by contrast, $\alpha < \alpha^*$, too few are in social groups, and the population in groups will be growing, with the "surplus" population leaving their group of origin and constituting new groups.

Finally, is the kind of sanctioning of norm violators which we have modelled commonplace in these societies? There is evidence that in some contemporary simple societies the lazy and the stingy are punished. Balikci (1970):177 reports the following concerning the Netsilik, an isolated tribe of Arctic hunters living on the Arctic coast:

> …there is a general rule…according to which all able bodied men should contribute to hunting, and the returns of the hunt should be shared according to established custom. Any activity in exception to this rule was bound to provoke criticism, various forms of conflict, and frequently social ostracism. (176)…lazy hunters were barely tolerated by the community. They were the objects of back biting and ostracism…until the opportunity came for an open quarrel. Stingy men who shared in a niggardly manner were treated similarly. (177)

And Lee (1979):458 reports that

> The most serious accusations one !Kung can level against another are the charge of stinginess and the charge of arrogance. To be stingy, or far-hearted, is to hoard one's goods jealously and secretively, guarding them "like a hyena." The corrective for this is to make the hoarder give "till it hurts"; that is to make him give generously and without stint until everyone can see that he is truly cleaned out. In order to ensure compliance with this cardinal rule the !Kung browbeat each other constantly to be more generous and not to hoard.

Lethal violence among the !Kung is quite high so the costs of these conflicts must sometimes be borne by those seeking to uphold norms of sharing (Lee 1979). By contrast to the reports of Lee and Balikci, however, Endicott (1988):118 reports horror expressed by a Batek informant at the thought of exiling a member whose laziness had caused some resentment.

More extensive evidence of punishment of norm violators is provided by Christopher Boehm's (1993) survey of the many studies in this area.

> ...intentional leveling linked to an egalitarian ethos is an immediate and probably an extremely widespread cause of human societies' failing to develop authoritative or coercive leadership. (226)

Bruce Knauft (1991):393,395 adds:

> In all ethnographically known simple societies, cooperative sharing of provisions is extended to mates, offspring, and many others within the band. ...This sharing takes place well outside the range of immediate kin, viz. among the diverse array of kin and non-kin who constitute the typical residence group of 25+ persons. Archeological evidence suggests that widespread networks facilitating diffuse access to and transfer of resources and information have been pronounced at least since the Upper Paleolithic...The strong internalization of a sharing ethic is in many respects the *sine qua non* of culture in these societies.

Using data from forty-eight surviving simple societies, Boehm (1993):228 concluded that

> the primary and most immediate cause of egalitarian behavior is a moralistic determination on the part of a local group's main political actors that no one of its members should be allowed to dominate the others.

Boehm further sought to determine whether intentional behavior (notably, social sanctioning) that had a leveling effect was widespread in such societies and more specifically whether it had any significant effects in suppressing the growth of authoritarian leadership. He found evidence that arrogant members of the group are constrained by public opinion, criticism and ridicule, disobedience, and extreme sanction:

> ...assassination is reported in 11 out of the 48....behaviors that terminated relations with an overly assertive individual or removed him from a leadership role involved 38 of the 48 societies, while in an additional 28 instances the person was manipulated

> by social pressure.... the great majority of these misbehaviors involve dominance or self-assertion. (231)
>
> among simple foragers, ... group execution of overassertive persons seems to be rather frequent. (239)

Other cases of costly enforcement of norms relevant to the model arise because its application is considerably more general than the case of working and shirking with which we have motivated it. Suitably emended, the model covers many generic cases of adherence to group-beneficial norms, and punishment for violation of these norms. The extension from team production to the sharing of food acquired individually has already been mentioned and is readily accomplished. A more ambitious extension is to the norm of monogamy, which if possible would considerably expand the scope of our model by encompassing what appears to be a quite common norm in hunter gather bands and a frequent occasion for the sanctioning of violators.

Suppose there is norm that restricts copulations to monogamous couples, which when violated leads to strife within a group or lessens its effectiveness in acquiring food, insuring against stochastic events, or defending itself, all of which reduce fitness levels of group members. Those who violate the norm, however, enhance their fitness by an amount $b$. Let $\sigma$ represent the fraction of those in the group violating the norm of monogamy, with $\sigma_r$ and $\sigma_n$ the fraction of reciprocators and non-reciprocators, respectively, violating the norm and suppose the group fitness costs of violations of the norm are simply linear in $\sigma$. In the absence of monitoring and ostracism, then, we have

$$\phi_n = q(1-\sigma) - b(1-\sigma_n)$$

$$\phi_r = q(1-\sigma) - b(1-\sigma_r),$$

where $q-b$ is just the fitness level in a group uniformly conforming to the norm with, as before $q > b$, so adherence to the norm is group beneficial. If we assume, as before that reciprocators are motivated both to observe the norm themselves ($\epsilon$) and to punish those who fail to observe it, we reproduce the working-shirking-monitoring model exactly. We are thus confident that the model as we have developed it is applicable to a wide range of concrete problems of norm adherence likely to arise in small stateless groups.

In sum, we think that the model, suitably extended to cover generic norm adherence and to accommodate movement between groups as well as group dissolution and formation, may adequately account for those fitness determining individual interactions in groups during the late Pleistocene.

We turn then to the second part of our question: does the model account for the proliferation of reciprocators in a population predominantly composed of non-reciprocators?

Such a population, a small fraction of whom are reciprocators, we will suppose, initially occupy positions in asocial groups, all experiencing the same level of fitness. If the many asocial groups are forming and dissolving by random draws from the population, one, by chance will have a distribution of types within the basin of attraction of $f_s$. It will then evolve as a region (d) social group with its equilibrium distribution of reciprocators. At this point we know that the members of this sole social group constitute a small fraction of the population so $\alpha < \alpha*$ and the average fitness of reciprocators, by 18, exceeds that of non-reciprocators, resulting in the growth of the population of the social group, which either sends migrants back to the asocial group or eventually divides. This process will continue until a sufficiently large number of social groups are in existence that at size $n$, their members constitute $\alpha^*$ of the population, at which point $d\tau/dt = 0$ and the population equilibrium we have described in Section 5 obtains.

## 7 Conclusion

We do not know that a human predisposition to strong reciprocity evolved as we have described. But it might well have. Our results convince us that an evolutionary process based on genetic inheritance under the influence of natural selection is capable of accounting for the considerable extent of strong reciprocity observed in contemporary society. If we are right, the experimental, historical and other evidence of strong reciprocity we introduced in Section 2 may appear to be expressions of human propensities rather than puzzling behaviors inviting *ad hoc* explanation.

Stronger conclusions would be premature, however, in view of the substantial further work to be done in this area. First, while there is ample evidence that genetic and cultural evolutionary processes may be jointly determined (Durham 1991, Feldman and Laland 1996, Aoki and Feldman 1997), we have not yet explored how a process of genetic evolution of traits underlying strong reciprocity might affect and be affected by a parallel process of cultural transmission of learned social behaviors, including the norm of punishing norm violators. Nor have we adequately explored the genetic evolutionary process we have modeled when stochastic variations in physical environments and migration flows generate endogenous variations in group population growth as well as occasional group dissolution. We thus do not know the stochastically stability properties of our various equilibria.[18] Moreover we have not addressed the adjustment speed and disequilibrium behavior of our model. Finally, we have not adequately explored how strong reciprocity might be represented formally in an otherwise conventional individual utility function.[19] We are convinced, however, that expanding the self-regarding outcome based preference framework to take account of these behaviors will enrich rather than vitiate the rational actor model of human behavior and better equip it to provide a basis for a unified social science.

## 8 Appendix A

Proof of Theorem 1: First, if $\mu$, the probability that a reciprocator monitors, is chosen to be a best response, we have

$$\mu \begin{cases} = 0, & c > \sigma p\rho \\ \in [0,1], & c = \sigma p\rho \\ = 1, & c < \sigma p\rho \end{cases} \tag{27}$$

Finally, if $\sigma_r$ and $\sigma_n$ are chosen as best responses to $\mu$, we have

$$\begin{cases} fn\mu ps < g_r & \sigma = 1 \\ g_r = fn\mu ps & \sigma_r \in [0,1],\ \sigma_n = 1 \\ g_r < fn\mu ps < g_n & \sigma_r = 0,\ \sigma_n = 1 \\ fn\mu ps = g_n & \sigma_r = 0,\ \sigma_n \in [0,1] \\ g_n < fn\mu ps & \sigma_r = \sigma_n = 0 \end{cases} \tag{28}$$

(a) For any $\mu \leq 1$ we have $fn\mu ps < g_r < g_n$, so $\sigma = \sigma_r = \sigma_n = 1$. But then $c < p\rho\sigma$ implies $\mu = 1$.

(b) Suppose first that $g_n/(nps + \epsilon) < f$. Since $fn\mu ps < g_n$, for any $\mu \leq 1$, we have $\sigma_n = 1$. Suppose $\mu = 0$. Then $g_r = g_n - \epsilon f > g_n - fnps > 0$, so $\sigma_r = 1$ But $\sigma_n = 1$, so $\sigma = 1$, so $\mu = 1$, a contradiction. Suppose $0 < \mu < 1$. Then $\sigma = c/p\rho$, so $\sigma_r$ is given by (5), which is negative, since $c/p\rho < 1 - f$. This is a contradiction, proving that $\mu = 1$. But then $g_r < fn\mu ps$, so $\sigma_r = 0$. Now suppose $g_n/\epsilon < f$. Then $\gamma_r < 0$ so $\sigma_r = 0$. Moreover $fnsp < g_n$, so $fnps\mu < g_n$, so $\sigma_n = 1$. Hence $\sigma = 1 - f$, which implies $\mu = 1$.

[18] On stochastic stability, see Kandori, Mailath and Rob (1993), Foster and Young (1990), Young (1993), and Young (1998).

[19] Rabin (1993), Levine (1996) and Falk and Fischbacher (1998) are promising starts in this direction.

(c) If $\mu = 1$, then $g_r < g_n < fn\mu ps$, so $\sigma = 0$, which implies $\mu = 0$, a contradiction. Suppose $\mu = 0$. Then if $f < g_n/\epsilon$, we have $g_r > 0$, so $\sigma_r = 1$, so $\sigma = 1$, so $\mu = 1$, a contradiction. If $f > g_n/\epsilon$, then $g_n < \epsilon f\mu = \epsilon f < fnps$, so $\sigma_n = 0$, so $\sigma = 0$, so $\mu = 0$, a contradiction. Thus $0 < \mu < 1$, so $\sigma = c/p\rho$. If $\sigma_r > 0$, then $\sigma_n = 1$ (if reciprocators are indifferent to working or shirking, or if reciprocators surely shirk, then non-reciprocators surely shirk). But then $c/p\rho = \sigma > (1-f)\sigma_n = 1-f$, which violates our assumption that $c/p\rho < 1-f$. Thus $\sigma_r = 0$, so $s_n = \sigma/(1-f) = c/p\rho(1-f)$.

(d) Note that $f < g_n/\epsilon$ implies $g_r > 0$. Suppose first that $fnps < g_n$. Then $fn\mu ps < g_n$, so for any $\mu \leq 1$, we have $\sigma_n = 1$. If $\mu = 1$, then $g_r < fn\mu ps$, so $\sigma_r = 0$. Then $\sigma = 1-f$, so $p\rho\sigma = p\rho(1-f) < c$, so $\mu = 0$, a contradiction. Hence $\mu < 1$. If $\mu = 0$, then $\sigma = 1$, since $g_r > 0$, so $p\rho\sigma > c$, so $\mu = 1$, a contradiction. Hence $0 < \mu < 1$, so $\sigma = c/p\rho$. Then $\sigma_r$ is given by (5), which is positive, since $c/p\rho > 1-f$.

Now suppose $g_n < fnps$. Then if $\mu = 1$, then $g_r < g_n < fn\mu ps$, so $\sigma = 0$, which implies $\mu = 0$, a contradiction. If $\mu = 0$, then $\sigma = 1$, since $g_r > 0$, so $\mu = 1$, a contradiction. Thus $0 < \mu < 1$, so $\sigma = c/p\rho$. If $\sigma_r = 0$ then $\sigma = (1-f)\sigma_n \leq 1-f < c/p\rho = \sigma$, a contradiction. Thus $\sigma_r > 0$, which implies, as in the previous paragraph, $\sigma_n = 1$, so $\sigma_r$ is given by (5), which is positive, less than unity. But then $\mu = g_r/fnps$.

(e) Since $f > g_n/\epsilon$, $g_r < 0$ so $\sigma_r = 0$. The cost of monitoring is nc and the expected gain satisfies

$$(1-f)np\rho\sigma_n \leq (1-f)np\rho < (c/p\rho)np\rho = cn.$$

Hence we must have $\mu = 0$. Thus $\sigma_n = 1$ and the rest follows. ■

## References

Abbink, Klaus, Gary E. Bolton, Abdolkarim Sadrieh, and Fang-Fang Tang, “Adaptive Learning versus Punishment in Ultimatum Bargaining,” 1996. Discussion Paper No. B0-381, University of Bonn.

Aoki, Kenichi and Marcus W. Feldman, “A Gene-Culture Coevolutionary Model for Brother-Sister Mating,” 1997. Working paper 97-05-046, Santa Fe Institute,.

Arrow, Kenneth J. and Frank Hahn, *General Competitive Analysis* (San Fransisco: Holden-Day, 1971).

— and Gerard Debreu, "Existence of an Equilibrium for a Competitive Economy," *Econometrica (*1954):265–290.

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

— and William D. Hamilton, "The Evolution of Cooperation," *Science* 211 (1981):1390–1396.

Balikci, Asen, *Netsilik Eskimo* (New York: Natural History Press, 1970).

Bergstrom, Theodore C., "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review* 85,1 (March 1995):58–81.

— and Oded Stark, "How Altruism can Prevail in an Evolutionary Environment," *American Economic Review* 83,2 (May 1993):149–155.

Blount, Sally, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior & Human Decision Processes* 63,2 (August 1995):131–144.

Blurton-Jones, Nicholas G., "Tolerated Theft, Suggestions about the Ecology and Evolution of Sharing, Hoarding, and Scrounging," *Social Science Information* 26,1 (1987):31–54.

Boehm, Christopher, "The Evolutionary Development of Morality as an Effect of Dominance Behavior and Conflict Interference," *Journal of Social and Biological Structures* 5 (1982):413–421.

— , *Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies* (Philadelphi, PA: University of Pennsylvania Press, 1984).

— , "Egalitarian Behavior and Reverse Dominance Hierarchy," *Current Anthropology* 34,3 (June 1993):227–254.

Boorman, Scott A. and Paul Levitt, *The Genetics of Altruism* (New York: Academic Press, 1980).

Boyd, Robert and J. Lorberbaum, "No Pure Strategy Is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game," *Nature* 327 (1987):58–59.

— and Peter J. Richerson, "The Evolution of Reciprocity in Sizable Groups," *Journal of Theoretical Biology* 132 (1988):337–356.

Breden, Felix, "Partioning of Covariance as Method for Studying Kin Selection," *Trends in Evolutionary Ecology* 5,7 (July 1990):224–228.

Cameron, Lisa, "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," 1995. Discussion Paper #345, Department of Economics, Princeton University.

Campbell, Donald T., "Two Distinct Routes beyond Kin Selection to Ultra-Sociality: Implications for the Humanities and Social Sciences," in Diane L. Bridgeman (ed.) *The Nature of Prosocial Development* (New York: Academic Press, 1983) pp. 11–41.

Cashdan, Elizabeth A., "Egalitarianism among Hunters and Gatherers," *American Anthropologist* 82 (1980):116–120.

Coleman, James S., "Free Riders and Zealots: The Role of Social Networks," *Sociological Theory* 6 (Spring 1988):52–57.

Crow, James F. and Motoo Kimura, *An Introduction to Population Genetic Theory* (New York: Harper & Row, 1970).

Damas, David, "Central Eskimo Systems of Food Sharing," *Ethnology* 11,3 (1972):220–240.

Davis, Douglas D. and Charles A. Holt, *Experimental Economics* (Princeton: Princeton University Press, 1993).

Dawkins, Richard, *The Selfish Gene, 2nd Edition* (Oxford: Oxford University Press, 1989).

Dugatkin, Lee Alan, *Cooperation among Animals* (New York: Oxford University Press, 1979).

Durham, William H., *Coevolution: Genes, Culture, and Human Diversity* (Stanford: Stanford University Press, 1991).

Endicott, Kirk, "Property, Power and Conflict among the Batek of Malaysia," in T. Ingold, D. Riches, and J. Woodburn (eds.) *Hunters and Gatherers* (New York: St. Martin's Press, 1988) pp. 110–127.

Eshel, I., "On the Neighbor Effect and the Evolution of Altruistic Traits," *Theoretical Population Biology* 3 (1972):258–277.

Falk, Armin and Urs Fischbacher, "Kindness is the Parent of Kindness: Modeling Reciprocity," 1998. Institute for Empirical Economic Research, University of Zürich.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation," 1997. Institute for Empirical Research in Economics, University of Zürich.

— and Simon Gächter, "Cooperation and Punishment," 1996. Working Paper, Institute for Empirical Economic Research, University of Zürich.

— , — , and Georg Kirchsteiger, "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica* 65,4 (July 1997):833–860.

Feldman, Marcus W. and Kevin N. Laland, "Gene-Culture Coevolutionary Theory," 1996. Working Paper No. 96-05-033, Santa Fe Institute.

Foley, Robert, *Another Unique Species: Patterns in Human Evolutionary Ecology* (New York: John Wiley and Sons, 1987).

Forsythe, Robert, Joel Horowitz, N. E. Savin, and Martin Sefton, "Replicability, Fairness and Pay in Experiments with Simple Bargaining Games," *Games and Economic Behavior* 6,3 (May 1994):347–369.

Foster, Dean and H. Peyton Young, "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology* 38 (1990):219–232.

Gintis, Herbert, *Game Theory Evolving* (Princeton, NJ: Princeton University Press, forthcoming).

Güth, Werner, "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory (*1995):323–344.

— and Menahem E. Yaari, "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in Ulrich Witt (ed.) *Explaining process and change: Approaches to evolutionary Economics* (Ann Arbor: University of Michigan Press, 1992) pp. 23–34.

— and Peter Ockenfels, "Efficiency by Trust in Fairness? Multiperiod Ultimatum Bargaining Experiments with an Increasing Cake," *International Journal of Game Theory* 22,1 (1993):51–73.

Güth, Werner, R. Schmittberger, and B. Schwarz, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3 (May 1982):367–388.

Guttman, Joel M., "Rational Actors, Tit-for-Tat Types, and the Evolution of Cooperation," *Journal of Economic Behavior and Organization* 29,1 (1996):27–56.

Hamilton, W. D., "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology* 37 (1964):1–16,17–52.

— , "Innate Social Aptitudes of Man: an Approach from Evolutionary Genetics," in Robin Fox (ed.) *Biosocial Anthropology* (New York: John Wiley and Sons, 1975) pp. 115–132.

Harpending, Henry and Alan Rogers, "On Wright's Mechanism for Intergroup Selection," *Journal of Theoretical Biology* 127 (1987):51–61.

Hawkes, Kristen, "Sharing and Collective Action," in E. Smith and B. Winterhalder (eds.) *Evolutionary Ecology and Human Behavior* (New York: Aldine, 1992) pp. 269–300.

— , "Why Hunter-Gatherers Work: An Ancient Version of the Problem of Public Goods," *Current Anthropology* 34,4 (1993):341–361.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith, "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology," April, 1996. Unpublished.

Kandori, M. G., G. Mailath, and R. Rob, "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica* 61 (1993):29–56.

Kaplan, Hillard and Kim Hill, "Food Sharing among Ache Foragers: Tests of EXplanatory Hypotheses," *Current Anthropology* 26,2 (1985):223–246.

— and — , "Hunting Ability and Reproductive Success among Male Ache Foragers: Preliminary Results," *Current Anthropology* 26,1 (1985):131–133.

— , — , Kristen Hawkes, and Ana Hurtado, "Food Sharing among Ache Hunter-Gatherers of Eastern Paraguay," *Current Anthropology* 25,1 (1984):113–115.

Kent, Susan, "And Justice for All: The Development of Political Centralization Among Newly Sedentary Foragers," *American Anthropologist* 93,1 (1989):703–712.

Klein, Richard G., *Human Career: Human Biological and Cultural* (Chicago: University of Chicago Press, 1989).

Knauft, Bruce, "Sociality versus Self-interest in Human Evolution," *Behavioral and Brain Sciences* 12,4 (1989):12–13.

— , "Violence and Sociality in Human Evolution," *Current Anthropology* 32,4 (August–October 1991):391–428.

Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory* 27 (1982):245–252.

Kropotkin, Petr, *Mutual Aid: A Factor in Evolution* (New York: Black Rose Books, 1989[1903]).

Lee, Richard Borshay, *The !Kung San: Men, Women and Work in a Foraging Society* (Cambridge: Cambridge University Press, 1979).

Levine, David K., "Modeling Altruism and Spitefulness in Experiments," 1996. Department of Economics, UCLA.

Maynard Smith, John, "Group Selection," *Quarterly Review of Biology* 51 (1976):277–283.

— , "The Origin of Altruism," *Nature* 393 (June 3 1998):639–640.

Moore, Jr. Barrington, *Injustice: The Social Bases of Obedience and Revolt* (White Plains: M. E. Sharpe, 1978).

Nisbett, Richard E. and Dov Cohen, *Culture of Honor: The Psychology of Violence in the South* (Boulder: Westview Press, 1996).

Ostrom, Elinor, James Walker, and Roy Gardner, "Covenants with and without a Sword: Self-Governance is Possible," *American Political Science Review* 86,2 (June 1992):404–417.

Rabin, Matthew, "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83,5 (1993):1281–1302.

Samuelson, Paul, "Complete Genetic Models for Altruism, Kin Selection, and Like-Gene Selection," *Journal of Social and Biological Structures* 6,1 (January 1983):3–15.

Scott, James C., *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia* (New Haven, CT: Yale University Press, 1976).

Sethi, Rajiv and E. Somanathan, "The Evolution of Social Norms in Common Property Resource Use," *American Economic Review* 86,4 (September 1996):766–788.

Sober, Elliot and David Sloan Wilson, *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998).

Trivers, R. L., "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971):35–57.

Uyenoyama, Marcy and Marcus Feldman, "Theories of Kin and Group Selection: A Population Genetics Approach," *Theoretical Population Biology* 17 (1980):380–414.

Wade, Michael, "Soft Selection, Hard Selection, Kin Selection and Group Selection," *American Naturalist* 125,1 (January 1985):61–73.

Walras, Leon, *Elements of Pure Economics* (London: George Allen and Unwin, 1954[1874]).

Wenzel, George W., "Ningiqtuq: Resource Sharing and Generalized Reciprocity in Clyde River, Nunavut," *Arctic Anthropology* 32,2 (1995):43–60.

Williams, G. C., *Adaptation and Natural Selection: A Critique of some Current Evolutionary Thought* (Princeton, NJ: Princeton University Press, 1966).

Wilson, David Sloan, *The Natural Selection of Populations and Communities* (Menlo Park, CA: Benjamin Cummings, 1980).

— and Elliott Sober, "Reintroducing Group Selection to the Human Behavioral Sciences," *Behavior and Brain Sciences* 17 (1994):585–654.

— and Lee A. Dugatkin, "Group Selection and Assortative Interactions," *American Naturalist* 149,2 (1997):336–351.

Wilson, Edward O., *Sociobiology: The New Synthesis* (Cambridge: Harvard University Press, 1975).

Woodburn, James, "Egalitarian Societies," *Man* 17,3 (1982):431–451.

— and Alan Barnard, "Property, Power and Ideology in Hunter-Gathering Societies: An Introduction," in T. Ingold, D. Riches, and J. Woodburn (eds.) *Hunters and Gatherers* (New York: St. Martin's Press, 1988) pp. 4–31.

Young, H. Peyton, "The Evolution of Conventions," *Econometrica* 61,1 (January 1993):57–84.

— , *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton, NJ: Princeton University Press, 1998).

e\Papers\Mutual Monitoring\Mutual Monitoring.tex July 30, 1998