The Evolution of Strong Reciprocity*

Samuel Bowles
413-548-9391 (Phone)
413-548-9852 (Fax)
bowles@econs.umass.edu
Department of Economics
University of Massachusetts, Amherst

Herbert Gintis
413-586-7756 (Phone)
413-586-6014 (Fax)
hgintis@mediaone.net
Department of Economics
University of Massachusetts, Amherst

April 14, 2000

*We would like to thank Christopher Boehm, Robert Boyd, Leda Cosmides, Steven Frank, Kristin Hawkes, Hillard Kaplan, Peter Richerson, Rajiv Sethi, Eric Alden Smith, E. Somanathan, Leigh Tesfatsion, and Peyton Young for their help with this paper, and the MacArthur Foundation for financial support.

The Evolution of Strong Reciprocity

Abstract

A number of outstanding puzzles in economics may be resolved by recognizing that where 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 motivated by self-regarding, outcome-oriented preferences. This behavior, which we call *strong reciprocity*, is a form of altruism in that it benefits others at the expense of the individual exhibiting it. While economists have doubted the evolutionary viability of altruistic preferences, we show that strong reciprocity can invade a population of non-reciprocators and can be sustained in a stable population equilibrium. Under assumptions that may reflect the relevant historical conditions, the model describes the genetic evolution of strong reciprocity as a component in the repertoire of human preferences.

1 Introduction

While the assumption of self interested action has proven a remarkably powerful behavioral foundation for economics, a number of important social phenomena are difficult to explain on this basis. Among these are the importance of fairness motives in wage setting and other exchanges involving strategic interaction (Blinder and Choi 1990, Bewley 1998), the extensive support for welfare programs even among those who cannot expect to be net beneficiaries (Gilens 1996, Luttmer 1998, Gilens 1999, Fong 2000, Piketty 1999) and the effectiveness of group incentives even where residual claimancy is shared among such a large number that the individual gains associated with one's own effort is vanishingly small (Ghemawat 1995, Hansen 1997, Knez and Simester 1998). Experiments by economists and psychologists have provided further evidence that in some situations nonselfish motives are robust predictors of behavior (Fehr and Gächter 2000, Fehr and Falk 1999, Kahneman, Knetch and Thaler 1986, Güth and Tietz 1990, Hoffman, McCabe and Smith 1998, Isaac, Walker and Williams 1994). While this evidence can be variously interpreted, we think it provides sufficient reason to consider a broader range of human motivations.

These studies suggest that where members of a group benefit from mutual adherence to a social norm, individuals may obey the norm and punish its violators, even when this behavior cannot be justified in terms of standard preferences. We call this *strong reciprocity*. We distinguish this from contingent cooperation in an indefinitely repeated game and other forms of mutually beneficial cooperation that can be accounted for in terms of self-interest. 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 (Fehr and Gächter 1999), from the ethnographic literature on simple societies (Knauft 1991, Boehm 1984, Boehm 1993), from historical accounts of collective action (Moore 1978, Scott 1976) , as well as from everyday observation.

Strong reciprocity confers group benefits by promoting cooperation and punishing free riding. However such behavior imposes individual costs, both because strong reciprocators contribute more to the group than selfish types, and because they sustain the costs of punishing free riders. 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.

We do not address the empirical question concerning the degree to which observed strongly reciprocal behavior is genetically as opposed to culturally based. Rather, we 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 and punishing, strong reciprocity can then evolve because of the greater likelihood that reciprocators will be in groups with effective mutual monitoring. This greater likelihood derives 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, we model a dynamical system in which genetically inherited preferences explain individual (not necessarily fitness-maximizing) behaviors, and population frequencies are governed by a fitness-based replicator dynamic derived from within-group interactions in a public goods game as well as migratory flows among groups.

We provide a population-level equilibrium in which strong reciprocity persists even though non-reciprocators have greater fitness when interacting with reciproca-

[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.

tors, and non-reciprocators may form a considerable fraction of the population (20% in a simulation we discuss below). We are also able to offer a plausible account of the successful invasion and diffusion of reciprocity behaviors in 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 similar to other accounts of reciprocity. The behaviors we seek to explain, while formally altruistic—that is individually costly and group beneficial—are more punishing than kind, a characteristic shared by Trivers (1971), Hirshleifer and Rasmusen (1989), Boyd and Richerson (1992), Sethi and Somanathan (1996) and Friedman and Singh (1999).[2] Like Binmore (1998) we use evidence on the evolution of humans in foraging bands to study the influence of reciprocity concerns on the nature of equilibria in public goods games, but unlike Binmore we explore the evolution of non-self-regarding preferences in these environments. And like Güth and Yaari (1992), Huck and Oechssler (1996), Bester and Güth (1998) and Friedman and Singh (1999), we distinguish between utility, which affects behavior, and fitness, which affects rates of reproduction.

Our approach is also distinctive in two respects. First, while most models of reciprocity use repeated interactions among pairs of agents to induce cooperative behavior (Boorman and Levitt 1980, Axelrod and Hamilton 1981, Kreps, Milgrom, Roberts and Wilson 1982, Axelrod 1984, Boyd and Lorberbaum 1987, Guttman 1996), we treat social interaction as 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.

Second, our model is based on group membership rather than genetic relatedness, as in Samuelson (1983), Bergstrom and Stark (1993), and Bergstrom (1995). However, unlike other models of this type (Robson 1990, Güth and Yaari 1992, Güth 1995) we do not assume reciprocators can be distinguished from non-reciprocators by some phenotypic trait, nor can individuals establish reputations by their behaviors. 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.

If our model is to account for the evolution of strong reciprocity in humans it should capture the social and physical environment of the foraging bands that made up most of human society for most of its history. While modern accounts of these societies record considerable variety in social organization and livelihood (Kelly

[2] Sethi and Somanathan's paper is most closely related with our work, but 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.

1995), the widespread sharing of food, valuable information, and other sources of survival among many of these societies in the modern world 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 working to acquire goods that later would be shared.

The evolutionary puzzle is not why group members work and share, but rather why they punish. To address this problem, we develop a team production model in which it is costly both to follow a work norm and to punish norm violators. Our model captures key of characteristics of small foraging bands.[3] 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—is 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) calls an "immediate return" system of production.

In Section 2 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 3 the rate of change of genetically different types within groups and in Section 4 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.

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

2 Equilibrium Working, Shirking and Punishing Within a Group

Consider a group with n members. Members may either work or shirk. If all shirk, they all have equal fitness ϕ_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 shrinks over time if its members all shirk. If there is no shirking, and if output is divided equally among members, each will have positive fitness.

However group members benefit from shirking while still sharing equally in the total production of the group. To model this, we suppose each member can either work, supplying one unit of effort, or shirk, supplying zero units of effort. Let σ_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.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.

The group now faces 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 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 ρ from punishing shirkers, 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, $b - f\epsilon$.[4] We assume both types are homogeneous. We assume throughout that $\rho p > c$, so the expected subjective benefits from punishing shirkers exceed

[4]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 $b - \epsilon$.

the cost of monitoring a shirker. We call the fitness-maximizers *non-reciprocators*. Finally, we assume that the utility of punishment accrues to reciprocators only if the level of cooperation in the group is strictly positive (this eliminates the uninteresting 'masochistic' equilibrium in which no one works and nevertheless reciprocators punish).

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 ρ and ϵ—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. 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 μ in equilibrium for all reciprocators, while non-reciprocators do not monitor. There will also be a common rate of shirking σ_r for reciprocators and σ_n for non-reciprocators, so if the proportion of reciprocators is f, we have

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

If a reciprocator monitors, the likelihood of detecting a non-reciprocator shirking is $\sigma_n p$, and the corresponding likelihood for a reciprocator is $\sigma_r p$ so the expected net cost of monitoring is[5]

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

For simplicity, we assume the probability that a shirker is detected not working when each of the reciprocators monitors at rate μ is linear in total monitoring, and

[5]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.

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 σ_n and σ_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 shall assume throughout that (a) $p\rho > c$, so that unless $\sigma = 0$, reciprocators monitor when the probability of detecting shirking is sufficiently high; (b) $b > q/n$, so that non-reciprocators will shirk if there is no punishment; and (c) $b < q/n + \epsilon$, so that in a group of all reciprocators there zero shirking even without punishment.

As described in Proposition 1 in the Appendix, this model has several possible equilibria, depending on the relationship between the fraction f of reciprocators and the various model parameters. To simplify the analysis, we shall assume the following inequality which, as shown in the Appendix, eliminates only implausible and uninteresting regions:

$$\frac{c}{p\rho} + \frac{g_n}{nps + \epsilon} > 1. \tag{5}$$

We then have

Theorem 1. *Under the stated assumptions the following cases are nonempty and exhaustive.*

(a) If $f < g_n/(nps + \epsilon)$ then $\sigma = \sigma_r = \sigma_n = 1$ and $\mu = 0$. This is the **asocial equilibrium** *in which all members shirk and there is no monitoring or punishment;*

(b) If $g_n/(nps+\epsilon) < f < g_n/\epsilon$ *then* $\sigma_n = 1$,

$$\sigma_r = \frac{1}{f}\left[\frac{c}{p\rho} - (1-f)\right], \tag{6}$$

so $\sigma = c/p\rho$ *and* $\mu = g_r/fnps$. *This is the* **social equilibrium** *in which non-reciprocators do not work and reciprocators have positive shirking, while monitoring with less than certainty;*

(c) If $g_n/\epsilon < f$, *then* $\sigma_n = 1$, $\sigma_r = 0$, $\sigma = 1-f$ *and* $\mu = 0$. *This is the* **unconditional cooperation equilibrium** *in which reciprocators never shirk and never monitor, while non-reciprocators shirk with certainty.*

This theorem follows directly from Proposition 1 in the Appendix, since the other cases mentioned in the theorem violate (5).

3 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 population dynamics within the group for the three 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 within-group fitness of the non-reciprocators. The reason is that while non-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 π_r and π_n be the rate of change of the reciprocators and non-reciprocators in a social group per time period taking account of the numbers lost through ostracism. Then if the fraction of reciprocators is f_t at time t, we have

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

Subtracting f_t, dividing by Δt, and passing to the limit, we get

$$\frac{df_t}{dt} = f_t(\pi_r - \pi) = f_t(1-f_t)(\pi_r - \pi_n). \tag{7}$$

Also, stability requires

$$\frac{d\pi_r}{df_t} < \frac{d\pi_n}{df_t}. \tag{8}$$

When f is in the social region, all agents receive fitness benefits $q(1-\sigma)$ as their share of group output while reciprocators bear a fitness cost of $b(1-\sigma)$ for working and μcn for monitoring. The fitness costs occasioned by the punishment of shirking, which are born by all non-reciprocators and σ_r of reciprocators, is $sfnp\mu$ which, using (4), we can express as g_r. The fitness of each type when in social groups is thus given by

$$\begin{aligned}\phi_n^s &= q(1-\sigma) - fnp\mu \\ \phi_r^s &= q(1-\sigma) - fnp\mu\sigma_r - b(1-\sigma_r) - \mu cn,\end{aligned}$$

and the fitness advantage of the non-reciprocator group members over the reciprocators is then $(\sigma_r - 1)g_r + b(1-\sigma_r) + \mu cn$. The expected contribution of each group member to the group's population in the next period is equal to their fitness minus the probability of ostracism if shirking, which is $fnp\mu = g_r/s$. Thus

$$\pi_n = \phi_n^s - fnp\mu, \tag{9}$$

$$\pi_r = \phi_r^s - \sigma_r fnp\mu. \tag{10}$$

Equating the two rates of increase, we find that the only equilibrium in the region is given by

$$f^* = \frac{g_n}{\epsilon} - \frac{bs(1-\sigma)}{\epsilon((1-\sigma)(1+s) - \sigma\rho)}. \tag{11}$$

Also

$$\frac{d}{df}(\pi_r - \pi_n)|_{f^*} = -\frac{\epsilon((1-\sigma)(1+s) - \sigma\rho)}{sf^*}. \tag{12}$$

Since $f^* < g_n/\epsilon$, the denominator in the second term in (11) must be positive. Therefore the numerator in (12) is positive, so the equilibrium is stable. In the unconditional cooperation region all members receive qf, but reciprocators pay b for working while non-reciprocators do not. Therefore $\pi_r - \pi_n = -b$ in this region, so the fraction f of reciprocators declines through the region. Figure 1 illustrate the within group dynamics and the stable equilibrium in the social region.

Figure 2 illustrates the stable equilibrium for a social group for a particular choice of the model's parameters $(c, \rho, s, b, q, n, p, \epsilon)$. The size of the group in the example and the other parameters have been deliberately chosen to make shirking highly fitness-beneficial for those who escape punishment and thus to illustrate a case in which both types are represented at substantial frequencies and shirking is not uncommon in the resulting equilibrium. The values are stationary in two senses.

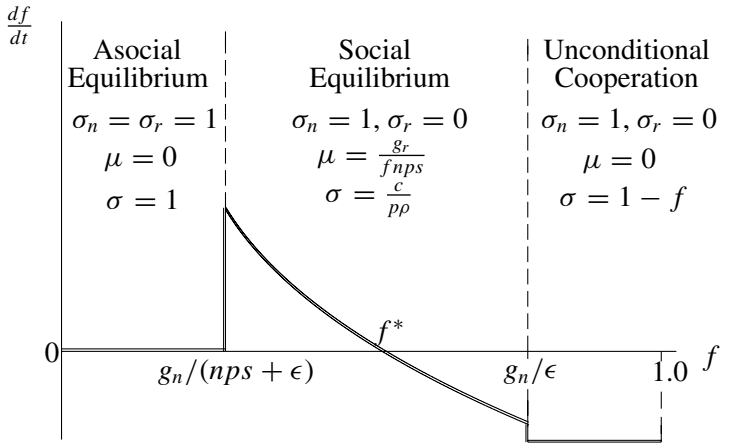

Figure 1: Within-Group Dynamics:$df/dt = \pi_r(f) - \pi_n(f)$, so $f = f^*$ is a stable equilibrium with basin of attraction $(g_n/(nps + \epsilon), 1]$, and $f = 0$ is neutrally stable for $f < g_n/(nps + \epsilon)$.

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 (12), the differential fitness of the non-reciprocators being offset by their greater likelihood of being ostracized.

4 Population Dynamics and Equilibrium

It remains to show that 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. To do this we first show that in equilibrium the composition of the social and asocial groups differ with reciprocators constituting a larger fraction in the former, and then show that population level average fitness of the two types is equalized at a positive value τ.

We assume the population consists of social and asocial groups. We assume all social groups are of size $n > 1$, and we ignore unconditional cooperation groups since they lie in the basin of attraction of the social groups. We assume social groups are in internal equilibrium as described in the previous section, with f^* being the fraction of those in social groups who are reciprocators, given by (11). We also define f_* to be the fraction of those in asocial groups who are reciprocators. We have $f_* < g_n/(nps+\epsilon)$, but the value of f_* is yet to be determined. We can express the fraction τ of reciprocators in the population as

$$\tau = \alpha f^* + (1 - \alpha) f_*, \tag{13}$$

Variable	Value	Description
μ	0.013	Monitoring Rate by Reciprocators
σ_r	0.501	Shirking Rate by Reciprocators
σ_n	1.000	Shirking Rate by Non-Reciprocators
σ	0.600	Average Shirking Rate
f^*	0.802	Frequency of Reciprocators
ϕ_r^s	0.095	Fitness of Reciprocators
ϕ_n^s	0.100	Fitness of Non-Reciprocators
$fnp\mu\sigma_r$	0.005	Rate of Ostracism of Reciprocators
$fnp\mu$	0.011	Rate of Ostracism of Non-Reciprocators
π_r	0.090	Rate of growth of Reciprocators in Group
π_n	0.090	Rate of growth of Non-Reciprocators in Group

Figure 2: Equilibrium Working, Shirking, and Punishing in the Social Region. The equilibrium for this region is generated using the following parameter values: $n = 40$, $q = 0.25$, $s = 0.0052$, $b = 0.011$, $p = 0.025$, $\rho = \epsilon = b/2$, $c = 0.000079$.

Let $-\beta$ be the immigration rate into social groups that maintains group size, so that

$$\beta = -\pi_r(f^*) = -\pi_n(f^*). \tag{14}$$

We assume members ostracized from a social group migrate to asocial groups. Also, if $\beta < 0$ (social group population is increasing), $|\beta|n$ 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). We shall assume the more plausible case that social groups are sufficiently fit that no post-ostracism immigration is need to maintain group size; i.e., $\beta \leq 0$.[6] Stationarity of f^* and f_* require that the composition of immigrants to the group be identical to the composition of those ostracised, the latter being just the ratio of the shirking probability of reciprocators to the average shirking rate in the social groups or

$$f_* = f^* \frac{\sigma_r}{\sigma} \tag{15}$$

(we prove this result in the Appendix). As $\sigma_r < \sigma$, (15) shows that the equilibrium frequency of reciprocators in asocial groups is less than their frequency in social groups.

[6]Our results continue to hold when $\beta > 0$ if β is not too large, but we believe that this case is implausible and do not consider it further.

Now let ϕ_r and ϕ_n be the average fitness of reciprocators and non-reciprocators in the population. We can then derive a replicator dynamic, as in (7), now defined at the population rather than the group level. The replicator equation for the fraction of reciprocators in the population is given by,

$$\frac{d\tau}{dt} = \tau(1-\tau)(\phi_r - \phi_n), \tag{16}$$

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 ϕ_r as follows. Let α_r be the fraction of reciprocators who are in social groups, then

$$\phi_r = \alpha_r \phi_r^s + (1-\alpha_r)\phi_a. \tag{17}$$

Similarly, if α_n is the fraction of non-reciprocators who are in social groups, we have

$$\phi_n = \alpha_n \phi_n^s + (1-\alpha_n)\phi_a. \tag{18}$$

Moreover if f^* and f_* are at their equilibrium values, α_r and α_n are determined by the distribution of the population between the social and asocial group, so we have

$$\alpha_r = \frac{\alpha f^*}{\alpha f^* + (1-\alpha) f_*}, \qquad \alpha_n = \frac{\alpha(1-f^*)}{\alpha(1-f^*) + (1-\alpha)(1-f_*)}, \tag{19}$$

where, as before, α is the fraction of the total population in social groups. The fitness of the two types in social groups is the number of members of the population contributed by each member of the group minus one or the post ostracism rate of growth of the group $(-\beta)$ plus the per person contribution to the population outside the group (by ostracism). Thus we have

$$\phi_r^s = \sigma_r \nu - \beta, \qquad \phi_n^s = \nu - \beta, \tag{20}$$

where $\nu = f^* np\mu$ is the rate at which shirkers are ostracized from social groups. From this it can be seen that the fitness of the reciprocators is lower: the two types contribute equally to the groups population (because f is stationary) while the non-reciprocators contribute more to the larger population.

Substituting these expressions in (17) and (18) and solving for the equilibrium condition $\phi_r = \phi_n$, we find that

$$\frac{\nu\sigma_r - \beta - \phi_a}{\nu\sigma - \beta - \phi_a} = \frac{\alpha_n}{\alpha_r}, \tag{21}$$

which requires that the relative fitness disadvantage of the reciprocators in social groups—the left side of (21)—be offset by the fitness disadvantage imposed on

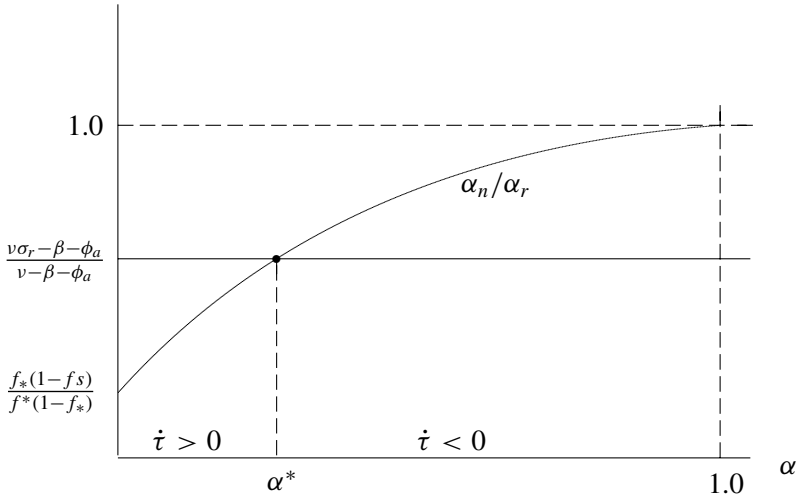

Figure 3: Population Level Equilibrium

non-reciprocators by their disproportionate location in asocial groups, given by the right side of (21).

The variables in (21) are completely determined by the parameter values underlying the within-group equilibrium, except for α_n/α_r, which rises monotonically in α, attaining a value of unity when all individuals are in social groups. Thus (21) determines the equilibrium fraction of the population in social groups, as is illustrated in Figure 3. 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 $1-(\nu\sigma_r-\beta-\phi_a)/(\nu\sigma-\beta-\phi_a)$ 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 α is

$$\alpha^* = \frac{-(\beta+\phi_a)}{\nu(1-f^*(1-s_r))-\beta-\phi_a}. \tag{22}$$

This expression is always less than unity, and is strictly positive if

$$\beta < -\phi_a. \tag{23}$$

The right hand side of (23) is strictly positive, so $\alpha^* \in (0,1)$ given our assumptions that $\beta \leq 0$ and $\phi_a < 0$.

Given α^*, the equilibrium distribution of types in the population is also determined, as the distribution of types within the asocial and social groups is unchanging. From (13) we have

$$\tau^* = \alpha^* f^* + (1-\alpha^*) f_*.$$

The dynamics of the population frequency of reciprocators is illustrated by Figure 3. 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. Because these results hold when $\alpha = 0$, a small group of mutant reciprocators would proliferate in a population of non-reciprocators.

Variable	Value	Description
τ	0.800	Population Frequency of Reciprocators
α	0.968	Fraction of Population in Social Groups
α_r	0.973	Fraction of Reciprocators in Social Groups
α_n	0.947	Fraction of Non-Reciprocators in Social Groups
f_*	0.670	Frequency of Reciprocators in Asocial Groups
ϕ_n	0.090	Average Fitness of Non-Reciprocators in Population
ϕ_r	0.090	Average Fitness of Reciprocators in Population
ϕ_a	-0.100	Fitness in Asocial Groups
ϕ_r^s	0.095	Fitness of Reciprocators in Social Groups
ϕ_n^s	0.100	Fitness of Non-Reciprocators in Social Groups

Figure 4: Reciprocators and Non-Reciprocators in Population Level Equilibrium. The parameter values used to generate the equilibrium are identical to those in the notes of Figure 2.

The population-level equilibrium supported by the parameter values used in the example of the within-group equilibrium (Figure 2) is shown in Figure 4. In this example most agents are in social groups, with the asocial status representing a temporary condition of those ostracized from social groups before relocating in a social group. Like the heterogeneity of the social groups and the high frequency of shirking illustrated in Figure 2, this aspect of our example may accurately reflect empirical realities in the relevant populations.

We then have

Theorem 2. *There is a unique equilibrium fraction of reciprocators $\tau^* > 0$. This equilibrium is stable in the replicator dynamic given by (16).*

Corollary 2.1. *A small number of reciprocators can invade a large population of non-reciprocators.*

To prove the theorem notice that $\alpha^* > 0$ implies $\tau^* > 0$ because social groups are unsustainable without reciprocators. But it is easily to show that $\alpha \in (0, 1)$,

because $\beta+\phi_a < 0$. Let us write $g(\alpha) = \alpha_n/\alpha_r$ and $g^* = (\nu\sigma_r-\beta-\phi_a)(\nu-\beta-\phi_a)$. Then $\tau^* = \alpha^* f^* + (1-\alpha^*) f_*$, where α^* is the solution to $g(\alpha) = g^*$. We have

$$g^* - g(0) = \frac{-(f^* - f_*)(\beta + \phi_a)}{(1 - f_*) f^* (\nu - \beta - \phi_a)},$$

which is strictly positive if $\beta + \phi_a < 0$. But if $g^* > g(0)$ then $d\tau/dt > 0$ by (16). Moreover $g(1) = 1 > g^*$, so by the mean value theorem there exists an $\alpha^* \in (0, 1)$ satisfying $g(\alpha^*) = g^*$. Also

$$g'(\alpha) = \frac{(f^* - f_*)(1 - f^*)}{f^*(1 - f_* + \alpha f_* - \alpha f^*)^2} > 0,$$

so $g(\alpha)$ is strictly increasing, implying that α^* is unique. Finally, since τ is a strictly increasing function of α, the equilibrium at α^* is clearly stable in (16). The corollary is true because $g^* > g(0)$. ■

5 The Evolution of Reciprocity

Can this model illuminate a process by which strong reciprocity might have become common in human populations? 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?[7] 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 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, we believe that our model may be illuminating.[8] 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

[7]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.

[8]Our main sources are listed in footnote 3. 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 and Kelly (1995).

> 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).[9] 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.

[9] 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.

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)

We have modelled punishment simply as ostracism from the group. But in the ethnographic record it takes several forms, including group fissioning to minimize interacting with shirkers and the withdrawal of cooperation from shirkers who remain co-resident. Extensions of the model to include these forms of punishment are straightforward. An excluded subgroup of shirkers, for example, would most likely have too few reciprocators to sustain the social equilibrium, and would simply become an asocial group, thus reproducing the effects of our individual level ostracism.

Our reading of the ethnographic and paleoanthropological evidence is that our model may capture the salient social and ecological conditions of the late Pleistocene. This alone is not adequate, of course, for we must also show that the model can account for the proliferation of reciprocators in a population composed of non-reciprocators, as our ancestral populations undoubtedly were.

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^*. It will then evolve as a 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 (21), 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 α^* of the population, at which point $d\tau/dt = 0$ and the population equilibrium we have described in Section 4 obtains.[10]

6 Conclusion

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 σ represent the fraction of those in the group violating the norm of monogamy, with σ_r and σ_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 σ. 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 be-

[10] We do not address the manner in which a small group of reciprocators might constitute a group and establish group norms except note that the process could easily come about simply by an extension to non-kin of common within-kin group practices (Boehm 1999).

fore that reciprocators are motivated both to observe the norm themselves (ϵ) 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.

We should stress, however, that any claim that strong reciprocity historically evolved by the mechanism we have identified remains entirely speculative. We are content to have shown that it could have. Moreover the mechanism underlying our model, while plausible, might be vulnerable to the emergence of actors who work ($\epsilon > 0$) but do not punish violators ($\rho = 0$). We do not regard this possibility as decisive for two reasons. First, the traits supporting adherence and punishment (ϵ and ρ) might be pleiotropically linked, the mutations effectively delinking the traits having either not occurred in this particular branch of hominids, or proven non-viable due to group extinctions among those experiencing these mutations, or for other reasons outside the model.[11] Second the cognitive and affective traits required to fashion, learn, detect violations of, and wish to uphold social norms may be genetically transmitted, while the content of the norms (and in particular the linking of ϵ and ρ) may be culturally transmitted. For example, one's unwillingness to join in the punishment of a norm violator (which according to Boehm (1993) is often collective and hence public) would itself be punished. Notice that this possible cultural linking of norm adherence and the punishment of violators does not trivialize the problem, as the fundamental puzzle remains, namely how could this individually costly mélange of behaviors overcome its fitness disadvantage within 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 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 may appear to be expressions of human propensities rather than puzzling behaviors inviting *ad hoc* explanation.

[11] Pleiotropic linking of traits is not merely a fortuitous possibility, but in fact is a likely evolutionary outcome in a situation where two traits separately are deleterious but together are fitness enhancing.

7 Appendix

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

(a) If $f < g_n/(nps + \epsilon)$ then $\sigma = \sigma_r = \sigma_n = 1$ and $\mu = 0$. This is the **asocial** *region members shirk and no member monitors.*

(b) If $f < g_n/nps$, $f < 1 - c/p\rho$, and $g_n/\epsilon < f$, 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 never shirk, but non-reciprocators work with positive probability, and reciprocators monitor with positive probability.

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

$$\sigma_r = \frac{1}{f}\left[\frac{c}{p\rho} - (1 - f)\right], \tag{24}$$

so $\sigma = c/p\rho$ and $\mu = g_r/fnps$. The is the **social** *region in which 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 $\mu = 0$. This is the **unconditional cooperation** *region, in which reciprocators never shirk and never monitor, while non-reciprocators shirk with certainty.*

Proof: First, if μ, the probability that a reciprocator monitors, is chosen to be a best response, we have

$$\mu \begin{cases} = 0, & c > \sigma p\rho \\ \in [0, 1], & c = \sigma p\rho \\ = 1, & c < \sigma p\rho \end{cases} \tag{25}$$

Finally, if σ_r and σ_n are chosen as best responses to μ, we have

$$\begin{cases} fn\mu ps < g_r & \sigma = 1 \\ g_r = fn\mu ps & \sigma_r \in [0, 1], \sigma_n = 1 \\ g_r < fn\mu ps < g_n & \sigma_r = 0, \sigma_n = 1 \\ fn\mu ps = g_n & \sigma_r = 0, \sigma_n \in [0, 1] \\ g_n < fn\mu ps & \sigma_r = \sigma_n = 0 \end{cases} \tag{26}$$

(a) For any $\mu \leq 1$ we have $fn\mu ps < g_r < g_n$, so $\sigma = \sigma_r = \sigma_n = 1$. But then $\mu = 0$ since by assumption reciprocators only monitor when there is positive working.

(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 σ_r is given by (24), 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 $\mu = 1$.

(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 σ_r is given by (24), 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 σ_r is given by (24), 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. ■

The intuition underlying the is illustrated by the depiction of the theorem's five cases in Figure 5. 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 (25) 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.

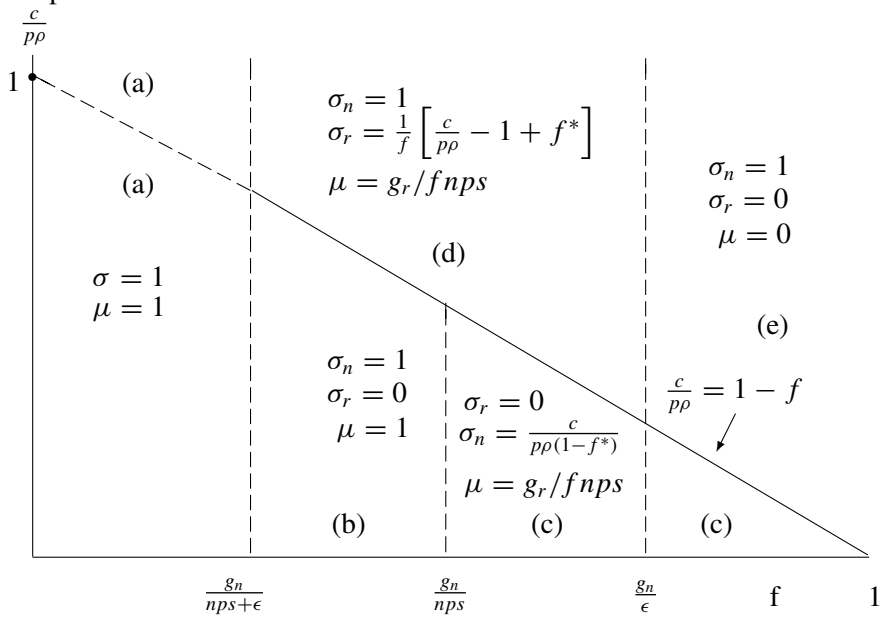

Figure 5: Case of Within-Group Interaction for Different Costs of Monitoring and Frequency of Reciprocators (cases (a) through (e) refer to the corresponding parts of Theorem 1). The figure assume $nps > \epsilon$, so one part of region (b) is not illustrated.

Figure 6 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 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 7 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.

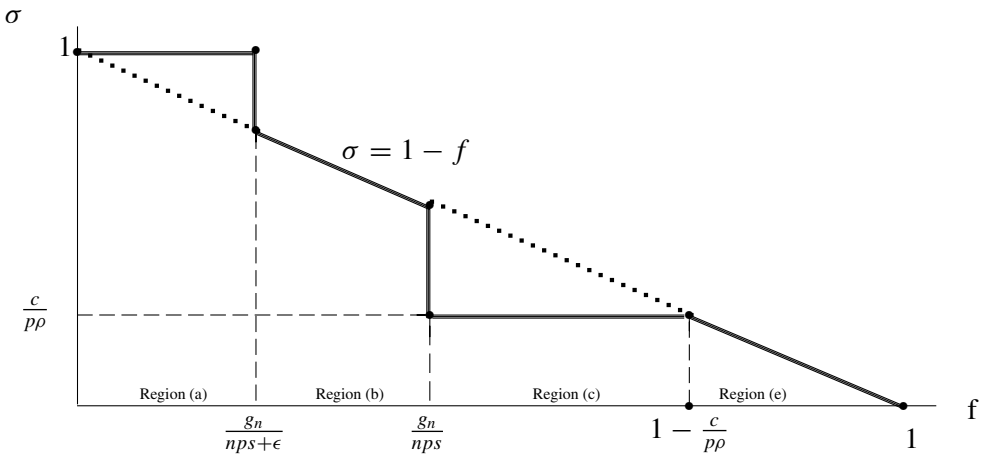

Figure 6: Relationship of Average Shirking to Fraction of Reciprocators with low monitoring costs ($c/p\rho < 1 - g_n/\epsilon$).

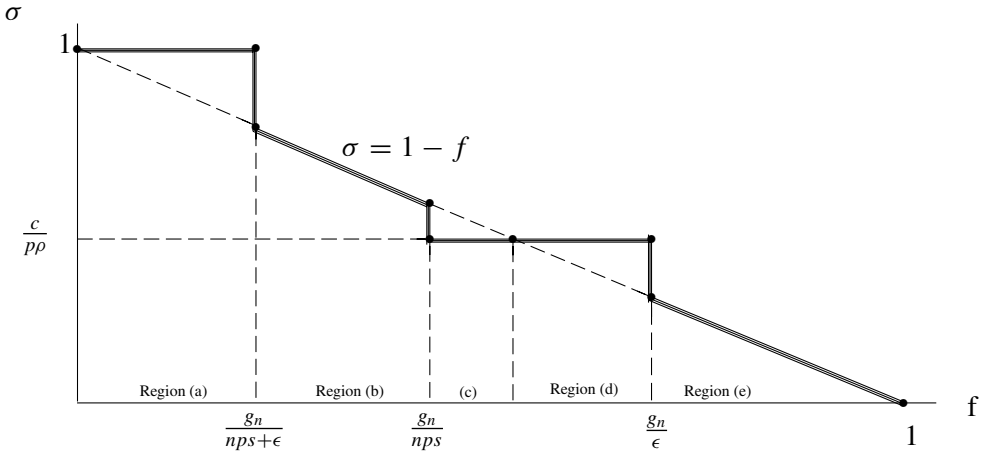

Figure 7: Relationship of Average Shirking to Fraction of Reciprocators with higher monitoring costs ($1 - g_n/\epsilon < c/p\rho < 1 - g_n/nps$).

Figure 8 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 $c/p\rho$.

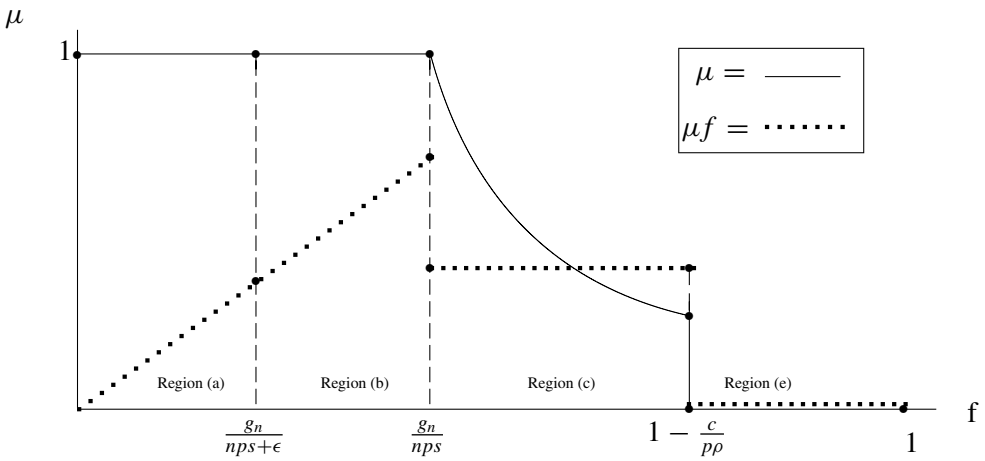

Figure 8: Monitoring Rate (μ) and total monitoring ($f\mu$).

Proof of Theorem 1: First, if μ, 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 σ_r and σ_n are chosen as best responses to μ, 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 σ_r is given by (24), 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 $\mu = 1$.

(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 σ_r is given by (24), 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 σ_r is given by (24), 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. ■

Proof of equation (15). For simplicity of exposition 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 ν be the rate at which shirkers are ostracized from social groups, and let σ_r be the rate at which reciprocators shirk in social groups (all

non-reciprocators shirk with certainty). Then if μ is the monitoring rate in social groups, using (11) and Theorem 1d, we have

$$\nu = f^* np\mu = \frac{g_r}{s} = \frac{b(1-\sigma)}{(1+s)(1-\sigma)-\rho\sigma} \tag{29}$$

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 α 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

$$f_{a,t}(1-\alpha)N(1+\phi_a\Delta t) + \alpha N(f^*\nu\sigma_r - \beta f_{a,t})\Delta t. \tag{30}$$

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

$$(1-\alpha)N(1+\phi_a\Delta t) + \alpha N(\nu\sigma - \beta)\Delta t. \tag{31}$$

Dividing (30) by (31), subtracting $f_{a,t}$, dividing by Δt and passing to the limit, we find that the fraction f_* of agents in asocial groups who are reciprocators satisfies the differential equation

$$\dot{f}_{a,t} = -\frac{\alpha\nu\sigma}{1-\alpha}(f_{a,t} - f_*), \tag{32}$$

where

$$f_* = f^*\frac{\sigma_r}{\sigma} \tag{33}$$

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.

We shall now prove that when the whole population is constant in size when in equilibrium, then we must have $\beta = 0$. We treat ϕ_r^s and ϕ_n^s as parameters, and solve (17), (18), (19) and (20) for the equilibrium condition $\phi_r = \phi_n$, getting

$$\alpha = \frac{f^*(\phi_r^s - \phi_a) + f_*(\phi_a - (f^*\phi_r^s + (1-f^*)\phi_n^s))}{(f^* - f_*)((f^*\phi_r^s + (1-f^*)\phi_n^s) - \phi_a)}. \tag{34}$$

The equation for zero total population growth is

$$\alpha(f^*\phi_r^s + (1 - f^*)\phi_n^s) + (1 - \alpha)\phi_a = 0, \tag{35}$$

which has solution

$$\alpha = \frac{-\phi_a}{f^*\phi_r^s + (1 - f^*)\phi_n^s - \phi_a}. \tag{36}$$

The condition for both (35) and (36) to hold is

$$f_* = \frac{f^*\phi_r^s}{f^*\phi_r^s + (1 - f^*)\phi_n^s}. \tag{37}$$

In equilibrium f_* satisfies (15) which, when substituted in (37) and simplified, give

$$\phi_r^s = \phi_n^s\sigma_r. \tag{38}$$

comparing this with (20), we see that $\beta = 0$.

References

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, *The 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.

Bester, Helmut and Werner Güth, "Is Altruism Evolutionarily Stable?," *Journal of Economic Behavior and Organization* 34,2 (February 1998):193–209.

Bewley, Truman F., "Why Not Cut Pay?," *European Economic Review* 42,3–5 (May 1998):459–490.

Binmore, Ken, *Game Theory and the Social Contract: Just Playing* (Cambridge, MA: MIT Press, 1998).

Blinder, Alan S. and Don H. Choi, "A Shred of Evidence on Theories of Wage Stickiness," *Quarterly Journal of Economics* 105,4 (November 1990):1003–15.

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* (Philadelphia, PA: University of Pennsylvania Press, 1984).

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

— , "Cooperation in Simple Societies," 1999. University of Southern California.

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, "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups," *Ethology and Sociobiology* 113 (1992):171–195.

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

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

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.

Fehr, Ernst and Armin Falk, "Wage Rigidity in a Competitive Incomplete Contract Market," *Journal of Political Economy* 107,1 (February 1999):106–134.

— and Simon Gächter, "Homo Reciprocans and Human Cooperation," 1999. University of Zurich.

— and — , "Cooperation and Punishment," *American Economic Review (*2000). forthcoming.

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

Fong, Christina, "Social Insurance or Conditional Generosity: The Role of Beliefs about Self- and Exogenous-Determination of Incomes in Redistributive Politics," 2000. Washington University Department of Political Science.

Friedman, Daniel and Nirvikar Singh, "On the Viability of Vengeance," 1999. Economics Department, UC Santa Cruz.

Ghemawat, Pankaj, "Competitive Advantage and Internal Organization: Nucor Revisited," *Journal of Economic and Management Strategy* 3,4 (winter 1995):685–

717.

Gilens, Martin, "'Race Coding' and White Opposition to Welfare," *American Political Science Review* 90,3 (September 1996):593–604.

— , *Why Americans Hate Welfare* (University of Chicago Press, 1999).

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.

Güth, Werner and Reinhard Tietz, "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results," *Journal of Economic Psychology* 11 (1990):417–449.

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.

Hansen, Daniel G., "Individual Responses to a Group Incentive," *Industrial and Labor Relations Review* 51,1 (October 1997):37–49.

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.

Hirshleifer, Jack and Eric Rasmusen, "Cooperation in a Repeated Prisoners' Dilemma with Ostracism," *Journal of Economic Behavior and Organization* 12 (1989):87–106.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith, "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology," *Economic Inquiry* 36,3 (July 1998):335–352.

Huck, Steffen and Jorg Oechssler, "The Indirect Evolutionary Approach to Explaining Fair Allocations," 1996. Humboldt University, forthcoming in *Games and Economic Behavior.*

Isaac, R. Mark, James M. Walker, and Arlington W. Williams, "Group Size and Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups," *Journal of Public Economics* 54 (May 1994):1–36.

Kahneman, Daniel, Jack L. Knetch, and Richard H. Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review* 76,4 (September 1986):728–741.

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.

Kelly, Robert L., *The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways* (Washington, DC: The Smithsonian Institution, 1995).

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.

Knez, Marc and Duncan Simester, "Firm-wide Incentives and Mutual Monitoring," September 1998. Graduate School of Business, University of Chicago.

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.

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

Luttmer, Erzo F. P., "Group Loyalty and the Taste for Redistribution," 1998. University of Chicago Business School.

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

Piketty, Thomas, "Attitudes Toward Income Inequaltiy in France: Do People Really Disagree?," 1999. CEPREMAP, Paris.

Robson, Arthur J., "Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake," *Journal of Theoretical Biology* 144 (1990):379–396.

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.

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

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.