DISCUSSION PAPER SERIES

IZA DP No. 1347

Social Reciprocity

Jeffrey P. Carpenter
Peter Hans Matthews

October 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Social Reciprocity

Jeffrey Carpenter
Middlebury College and IZA Bonn

Peter Hans Matthews
Middlebury College

Discussion Paper No. 1347
October 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

IZA Discussion Paper No. 1347
October 2004

ABSTRACT

Social Reciprocity

We define social reciprocity as the act of demonstrating one's disapproval, at some personal cost, for the violation of widely-held norms (e.g., don't free ride). Social reciprocity differs from standard notions of reciprocity because social reciprocators intervene whenever a norm is violated and do not condition intervention on potential future payoffs, revenge, or altruism. Instead, we posit that social reciprocity is a triggered normative response. Our experiment confirms the existence of social reciprocity and demonstrates that more socially efficient outcomes arise when reciprocity can be expressed socially. To provide theoretical foundations for social reciprocity, we show that generalized punishment norms survive in one of the two stable equilibria of an evolutionary game with selection drift.

JEL Classification: C92, Z13

Keywords: reciprocity, norm, experiment, public good, learning, evolution

Corresponding author:

Jeffrey Carpenter
Department of Economics
Middlebury College
Middlebury, VT 05753
USA
Email: jpc@middlebury.edu

1 Introduction[1]

Four decades have passed since the infamous murder of Kitty Genovese in Queens, New York, in 1964, but for those who lament the state of urban life in the United States then and now, her name still resonates. It is our view, however, that what most shocks us about the crime is not the appearance of widespread indifference to the pain of others, but rather that such indifference is still the exception, not the rule, despite standard assumptions about the character of *homo economicus*. In particular, we suspect that our desire to "punish" in this case is not limited to her murderer but extends, albeit in a different sense and to a smaller extent, to her neighbors.

The subsequent research of psychologists and sociologists on "bystander intervention," much of it motivated by the Genovese case, provides some support for this interpretation. Borofsky et al (1971) and Shotland and Straw (1976), for example, demonstrated that a significant number of people will intervene in a seemingly severe altercation between two people even though the one to intervene is not being harmed, nor is there any reason to expect that the one to intervene will receive any payoff from doing so. In the former, 29 percent intervened in situations in which two confederates of the experimenter staged an altercation that escalated into a physical fight. The latter found that a much higher proportion, 65 percent, intervened when a male confederate pretended to assault a female confederate, but also that this number dwindled to 19 percent when the two confederates seemed to be married. (The difference was attributed to differences in the costs of intervention: Shotland and Straw concluded that bystanders believed that husbands were more likely to stay and fight.)

In economic environments, intervention and punishment are often important in social dilemmas, for example, the provision of local public goods, common pool resource extraction or team production. Despite strong incentives to free ride on the efforts of others, the members of groups who confront such dilemmas are sometimes adept at attenuating incentive problems without external intervention. Communities often develop rules that make contributing and free-

[1] We thank Marco Castillo, Jeremy Clark, David Colander, Carolyn Craven, Herb Gintis, Corinna Noelke, Louis Putterman and David Sloan Wilson for their comments on previous drafts, as well as seminar and conference participants at the European University Institute, the Economic Science Association, the Canadian Economics Association and Middlebury College for their constructive feedback. We also thank Okomboli Ong'ong'a for research assistance, and Middlebury College and the National Science Foundation (SES-CAREER 0092953) for financial support.

riding transparent (Ostrom, 1992), but perhaps more importantly community members are also often willing to incur costs to monitor and punish behavior that benefits the individual but harms the group (e.g., Acheson, 1988). Acts of this kind tend to maintain or increase the efficiency of social interactions so one might posit that monitoring is in the interests of group members. Indeed, it may be the case that if free riders respond by contributing more in the future, the benefits that accrue to monitoring and punishing may exceed the individual costs, measured perhaps in terms of possible retaliation. However, even in this case we know that punishers can do better by free riding on the punishment meted out by other monitors. In fact, by the same logic that not contributing dominates contributing to a public good in one-shot interactions (Olson, 1965), there is no logic by which narrowly self-interested agents monitor and punish.

A number of solutions to this paradox have been offered. Axelrod's (1984) famous tournament, for example, illustrates what can happen when interactions are repeated. If there is some chance that the interaction will continue for another period and if those involved are not too impatient, strategies that punish (tit-for-tat, for example) can support Nash equilibria in which individuals cooperate and punish those who do not. (This is of course the intuition behind the so-called Folk Theorem.) One cannot, however, explain cooperation or punishment in one-shot interactions on this basis, and the proposition that these are simply "mistakes" by individuals who believed otherwise is difficult to rationalize in the context of the systematic behavior observed in experiments. Fehr and Fischbacher (2001) conclude, in fact, that even a naïve decision maker will find the difference between one-shot and repeated encounters a salient one.

Other researchers have considered alternative representations of preferences and the influence of social norms. For example, Kandal and Lazear (1992) show that contributions to team production can be sustained at considerable levels if team members are motivated by peer pressure. Altruists may also punish free riders because they want to increase the payoffs of the other, contributing, group members. Reciprocity may also cause players to retaliate against free riders, either because their cooperation has been exploited and/or the lower levels of public goods provision impose material costs (Bowles et al 2001). However, reciprocity (or conditional cooperation) by itself is unable to maintain cooperation without punishment because reciprocators have only one way to show their unhappiness with free riders - they withhold contributions themselves. This implies that even a small amount of free riding can ruin group-level cooperation (Fehr and Fischbacher, 2003).

In this paper we are interested in understanding the origins, limits, and social implications of individuals who incur costs to express their disapproval of antisocial behavior. Our focus is on norm-driven reciprocity and, in particular, on the willingness of individuals to punish such behavior *both* when the punisher him/herself has been harmed and when neither the punisher, nor his/her group, has been harmed. Little or no attention has been paid to the latter, a form of "third party punishment," in the economics literature, with the notable exception of Fehr and Fischbacher (2004). Fehr and Fischbacher find strong evidence of third party punishment in their three-person dictator experiment, but suggest that our public goods-oriented design "allows for reciprocity and strategic interactions among third parties ... [so that we] cannot rule out third party punishment for reasons of self-interest." Under our protocol (Appendix A), however, participants never knew who had, or had not, punished whom, so we are confident that self-interest is *not* the explanation.

To motivate these two very different scenarios, we distinguish between two types of norm-driven reciprocal behavior based on group boundaries. *Strong Reciprocators* (Bowles and Gintis, 2003; Carpenter et al., 2004; Gintis, 2000; Sethi, 1996) punish those members of their ingroup that free-ride, where an ingroup is loosely defined as the subset of individuals who benefit from a specific public good that they can all contribute to. *Social Reciprocators*, on the other hand, punish free-riders even in groups to which they can neither contribute to nor benefit directly from. Social reciprocity differs from strong reciprocity because social reciprocators punish all norm violators, regardless of group affiliation and with little regard to the social distance between punisher and norm violator, as long as there exists some "punishment network" that connects them. Further, while the trigger for punishment by strong reciprocators is the cost implicitly imposed by a free-rider on the group, we hypothesize that the trigger for social reciprocity is simpler. Social reciprocators just punish anyone who violates a contribution norm, and need not be harmed directly by the free-rider.

One could also frame the relationship between strong and social reciprocity in terms of "fuzzy boundaries": social reciprocity is the natural extension of strong reciprocity when group boundaries are not sharp. Urban neighborhoods are a classic example of the fuzzy boundary: it is often not obvious where one neighborhood starts and another ends. Another example occurs in team production when multiple teams occupy the same shop floor. In this situation, strong reciprocity dictates that the members of a specific term punish the shirkers on that team and no others. By contrast, social reciprocity requires them to sanction all shirkers on all teams.

The psychological experiments on bystander intervention mentioned above offer two more examples of social reciprocity, and a third can be found in Latane and Darley's (1970) work. In their experiment, subjects are asked to wait in a room to be interviewed. A confederate, also in the room, steals what remains of the show-up fee fund when the experimenter leaves. Their dependent variable is the probability that subjects report the theft when the experimenter returns. Because all subjects have been paid their show-up fee and therefore suffer no loss when the theft occurs, strong reciprocity is not an issue. Furthermore, since there is no expectation of a reward, there can be no instrumental reason for intervening so, not least because the costs of turning in the confederate could be high. Despite this, in 50% of the cases in which the subjects reported noticing the theft, they turned in the confederate.

Identifying and understanding socially reciprocal behavioral types that indiscriminately punish deviations from widely held norms is important because societies in which such behavior is present will be more cooperative, provide public goods at higher levels, be better able to complete contracts in information-poor environments, and extract from common pool resources more conscientiously than both non-reciprocal societies and societies based on standard notions of reciprocity alone. Provided free riders react to punishment by contributing more and fulfilling commitments, societies in which people punish all rule breakers do better because antisocial behavior will be detected more often and punished more severely.

To develop the case for social reciprocity, we proceed as follows. In the next section we present a summary of the existing evidence supporting the role of reciprocity-based monitoring regimes in both field settings and in the experimental lab. Sections 3 through 7 outline the design and results of an experiment we conducted to test for social reciprocity in an environment where it is costly to punish. In the penultimate section, we then provide some theoretical foundations for social reciprocity by showing that agents who punish outgroup norm violators survive in one of two stable equilibria of an evolutionary public goods game with drift. This section is important, not only because it provides reasonable microfoundations for the social phenomenon we are interested in, but also because the model allows us to provide an integrated approach to the topic. We discuss the implications of our results in Section 9.

2 The Existence of Reciprocity-Based Monitoring Schemes

In this section we summarize the existing evidence that suggests people, facing social dilemmas, engage in peer monitoring. We will consider evidence from both experiments and field studies. While the experiments we discuss were designed only to test for peer monitoring within specific groups, our examples from the field suggest that monitoring may transgress group boundaries. This fact provides the impetus for studying social reciprocity directly.

Peer monitoring has been tested experimentally in two specific game environments, common pool resource experiments where participants contribute by showing restraint when extracting from a commons and voluntary contribution experiments in which participants decide whether or not to contribute to a public good, the benefits of which are shared by the entire group. Ostrom et al. (1992), using a common pool resource design, were the first to demonstrate efficiency gains from peer monitoring. Their results showed that participants were able to sustain significant efficiency gains when they were allowed to punish those who extracted too much from the commons. These findings were later extended in Ostrom et al. (1994) and replicated in Moir (1998).

The first public goods experiment incorporating peer monitoring was conducted by Fehr and Gächter (2000) who confirmed the reciprocity-based theory of play in public goods games originating in Andreoni (1988). Andreoni's experimental design is noteworthy because it was able to differentiate learning from reciprocity. More specifically, the design had participants play a multi-period voluntary contribution game twice in a row (without knowing there would be a second game). The first play of the game resulted in the standard decay of contributions which might suggest that players learned to free ride. However, instead of starting at low levels of contributions, the second play began with contributions significantly higher than at the end of the first play suggesting that, rather than learning to free ride, participants withheld contributions in the first play to get back at free riders. When allowed to directly punish the other group members, Fehr and Gächter showed that free riders are punished and contributions do not decline.

The work of Fehr and Gächter has subsequently been replicated and extended in a number of interesting directions. Bowles et al. (2001) develop a reciprocity-based model of team production which predicts punishment in equilibrium and tests the model experimentally. Their results indicate that the propensity to punish a shirking team member is directly proportional to how much harm the shirker inflicts on the punisher and that shirkers respond to punishment by contributing more in the future. Additionally, Carpenter (2004) shows the effectiveness of peer monitoring need not be attenuated in large groups. Page and Putterman (2000) also confirm that punishment is used to maintain or increase contributions to a public good and show that communication among players, which usually increases contributions, has mixed effects when combined with sanctions. Finally, Sefton et al. (2000) ran an experiment in which players could reward and sanction other players. When both rewards and sanctions are allowed, they show that initially, rewards are used, but by the end of the experiment rewards abate and players rely mainly on sanctions.

Summarizing the results of previous experiments, we see that peer monitoring occurs and can be explained by the existence of reciprocally-motivated players who punish players who inflict costs on them (e.g. reduced payoffs from the public good) by free riding.

Although the evidence is less direct than that generated in the experimental lab, field studies of common pool resources, team production, and on a larger scale, neighborhoods also suggest that free riding and antisocial behavior can be controlled by peer monitoring. For example, Acheson (1993) illustrates how members of small, local fisheries prevent over-extraction by relying on endogenously evolved norms (that are often illegal) to punish over-extractors. Likewise, the Craig and Pencavel (1995) study of plywood cooperatives and the Ghemawat (1995) paper on a steel mini mill show that productive teams control shirking endogenously without the need of supervisors. Lastly, Sampson et al. (1997) show that, controlling for previous violence and individual characteristics, community monitoring, which they term collective efficacy, can explain differences in the amount of antisocial behavior occurring in different neighborhoods of Chicago. In short, case and field studies of actual social dilemmas indicate that groups regulate free riding endogenously and, given existing experimental results, the most parsimonious explanations are reciprocity-based.

The study of Sampson et al. is particularly interesting to us because neighborhoods are often populated with relatively large groups and are often distinguished by fuzzy borders while fisheries and work teams are generally smaller and more well-defined. It follows that egoistic incentives to monitor in neighborhoods are low because the benefits of monitoring are diffuse. This phenomenon suggests that monitoring free riders and community policing, in general, transgress blurry group boundaries. Therefore, the apparent efficiency of selected communities can not be explained by egoistic reasons to punish free riders or narrowly defined notions of reciprocity based on the intimacies of small groups in which reciprocators punish transgressors who impose costs on them directly.

3 A Social Reciprocity Experiment

We designed a public goods experiment to test for the existence of social reciprocity and to differentiate it from other theories of punishment (i.e. strong reciprocity and altruism). While our design is based on the standard voluntary contribution mechanism originally used in Isaac et al. (1984), to test whether players will punish free riders we allow players to monitor the decisions made by other players and punish them at a cost. To differentiate social reciprocity from other punishment explanations we developed additional design features that provided a game environment in which only players who don't respond to the material costs imposed on them would punish a subset of free riders. The specifics of our experiment are as follows.

We recruited ninety-six participants (thirty-five percent were female) in eleven experimental sessions. The participants were assigned to twenty-four four-person groups and each participant remained in the same group for all ten periods of the experiment. The fact that the game lasted only ten periods was common knowledge. Participants earned an average of $16.55 including a $5 show-up fee and a typical session lasted slightly less than an hour.

There were three treatments: a replication of the standard voluntary contribution game (VCM) which we use as a control on our procedures (4 groups), a replication of previous peer monitoring experiments in which players could monitor and sanction other members of their group (6 groups), and our social reciprocity treatment in which players could monitor and punish all the other players in a session, but they only benefited from their own group's contribution to a public good (14 groups).

The payoff function for the social reciprocity treatment was similar to the mutual monitoring incentive structure (see Bowles et al., 2001), but we augmented it to account for what we will call outgroup punishment. *Outgroup punishment* occurs when a member of one group sanctions a member of another group. Likewise, *ingroup punishment* occurs when members of a group punish each other. In the VCM treatment no punishment was allowed. In the strong reciprocity treatment no outgroup punishment was allowed and players saw only the contributions of their group members. But, in the social reciprocity treatment participants saw the contributions of all players and could punish any other participant in the session. Punishment was costly; players paid one experimental monetary unit (EMU) to reduce the gross earnings of another player by two EMUs.[2]

Imagine n players divided equally into k groups, each of whom can contribute any fraction of their w EMU endowment to a public good, keeping the rest. Say player i in group k free rides at rate $0 < \sigma_i^k < 1$ and contributes $(1 - \sigma_i^k)w$ to the public good, the benefits of which are shared only by members of group k. Each player's contribution is revealed to all the other players in the session, who then can punish any other player at a cost of 1 EMU per sanction. Let s_{ij} be the expenditure on sanctions assigned by player i to player j (we force $s_{ii} = 0$). Then the payoff to player i in group k is:

$$\pi_i^k = [\sigma_i^k + (n/p)m(1 - \sigma^k)]w - \sum s_{ij} - 2\sum s_{ji}$$

where $\sigma^k \equiv \left(\sum \sigma_i^k\right)/n$ is the average free riding rate in group k, $\sum s_{ij}$ is player i's expenditure on sanctions and $2\sum s_{ji}$ is the reduction in i's payoff due to the total sanctions received from the rest of the players. The variable m is the marginal per capita return on a contribution to the public good (see Ledyard, 1995). In all sessions m was set to 0.5 and w was set to 25 EMUs.

With $m = 0.5$, the dominant strategy is to free ride on the contributions of the rest of one's group (i.e. $\sigma_i^k = 1$ for all i) because each contributed EMU returns only 0.5 to the contributor. Also notice that if everyone in a four-person group contributes one EMU, they all receive a return of 2 EMUs from the public good. Therefore, these incentives form a social dilemma - group incentives are at odds with individual incentives. Considering punishment, because sanctions are costly to impose and their benefit cannot be fully internalized (ingroup) or cannot be internalized at all (outgroup) by the punisher, it is incredible and therefore cannot be a component of any subgame perfect equilibrium. Because punishment is an incredible threat, no one should fear it and therefore the only subgame perfect equilibrium in this game is where everyone free rides and nobody punishes. We feel, these incentives provide a stringent test of social reciprocity. In this environment social reciprocity is expressed when players punish

[2] The instructions referred to "reductions" with no interpretation supplied.

free riders outside their groups. Outgroup punishment can not be explained by strong reciprocity because free riders in other groups inflict no harm on the punisher. Outgroup punishment can also not be explained by tit-for-tat because there are no possible future benefits.

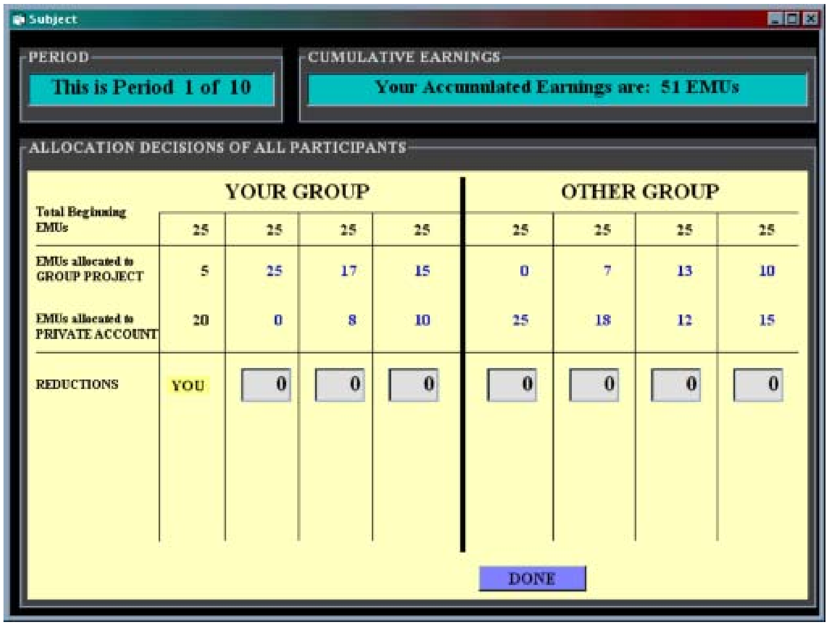
In the social reciprocity treatment each session was composed of two separate groups playing simultaneously. A session lasted ten periods and each period had three stages which proceeded as follows.[3] In stage one players contributed any fraction of their 25 EMU endowment in whole EMUs to the public good. The group total contribution was calculated and reported to each player along with his or her gross payoff for the period. Participants were then shown the contribution decisions of all the other players in the session. Figure 1 is a screen shot of what participants saw at the second stage. Players imposed sanctions by typing the number of EMUs they wished to spend to punish an individual in the textbox below that player's decision. After all players were done distributing sanctions, the experiment moved to stage three where everyone was shown an itemized summary of their net payoff (gross payoff minus punishment dealt minus punishment received) for the period. However, it is important to note that players never knew where the punishment that they received came from. Specifically, they never knew which individual or set of individuals punished them, nor did they know from which group punishment originated. This anonymity of punishment is important because it prevents two phenomena that could confound our results. First, anonymity prevents punishment feuds between individuals within a group or between groups and second, anonymity prevents between-group reciprocity from arising.

4 Does Social Reciprocity Exist?

The first question we wish to address is whether our participants (or a significant fraction of them) exhibit social reciprocity. Similar to other studies of punishment in social dilemma games, an overwhelming majority of our participants punished. Specifically, 82% of our subjects sanctioned ingroup and 50% punished outgroup at least once. Hence, a preliminary look at our data suggests half our participants exhibit some degree of social reciprocity.

Figure 2 presents a summary of contributions in our three treatments. The vertical axis measures the fraction of the individual endowment (25 EMUs)

[3] The participant instructions are provided in Appendix A.



PERIOD

This is Period 1 of 10

CUMULATIVE EARNINGS

Your Accumulated Earnings are: 51 EMUs

ALLOCATION DECISIONS OF ALL PARTICIPANTS

	YOUR GROUP				OTHER GROUP			
Total Beginning EMUs	25	25	25	25	25	25	25	25
EMUs allocated to GROUP PROJECT	5	25	17	15	0	7	13	10
EMUs allocated to PRIVATE ACCOUNT	20	0	8	10	25	18	12	15
REDUCTIONS	YOU	0	0	0	0	0	0	0

Figure 1: The Social Reciprocity Treatment Punishment Screen Shot.

contributed to the public good, on average. As one can see, our baseline, VCM treatment replicates the standard decline in contributions seen in many public goods experiments (see Ledyard, 1995 for a survey). This implies there is nothing strange about our protocol or subject pool. We also see that peer monitoring (i.e., restricting players to ingroup punishment only) largely maintains the initial level of cooperation. This behavior is consistent with prior peer monitoring experiments (see Bowles et al., 2001; Page and Putterman, 2000; and Sefton et al., 2000). Interestingly, and confirming our prior concerning the implications of social reciprocity, contributions are highest when players can punish free riders both inside and outside their groups. Further, these contribution differences are all significant at better than the 99% level.[4] However, there appears to be an end-game effect in contributions. Contributions drop substantially from round eight to round ten in both punishment treatments, but players in the

[4] We assess this by regressing group total contributions on treatment indicator variables and accounting for the upper and lower limits of contributions using the Tobit procedure and individual group heterogeneity by including random effects. The resulting estimate is: $Cont_{group} = 51.07 + 24.75Social + 15.05Strong$ and both coefficients are significant at the 99% level. Lastly, to complete the comparisons, the two point estimates are also highly significantly different ($p < 0.01$).

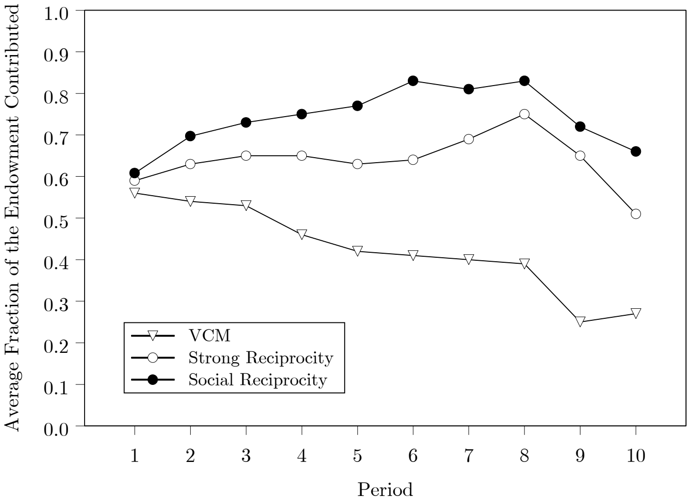


Figure 2: Average Contributions (VCM is the standard voluntary contribution mechanism, 4 groups; Strong Reciprocity is where only ingroup punishment is allowed, 6 groups; and Social Reciprocity is where players can punish both ingroup and outgroup, 14 groups).

social reciprocity treatment react less to the endgame. Despite the end-game effect, our first major result is that social reciprocity exists and is associated with increased contributions to a public good.

Concerning punishment expenditures, the first thing to notice in Figure 3 is that our strong reciprocity treatment seems to elicit more ingroup punishment than the social reciprocity treatment. However, one should be careful drawing this conclusion because, as was just mentioned, contributions are significantly higher in the social reciprocity treatment which means less punishment was warranted. Our second observation is that within the social reciprocity treatment it appears players spend more resources punishing ingroup than outgroup players. However, while this appears to be the case when looking at Figure 3, the pooled average difference between ingroup and outgroup sanctions (including all those cases when no punishment was levied) is not highly significant, t=-2.15, p=0.03 and the two types of punishment are not distributed differently,

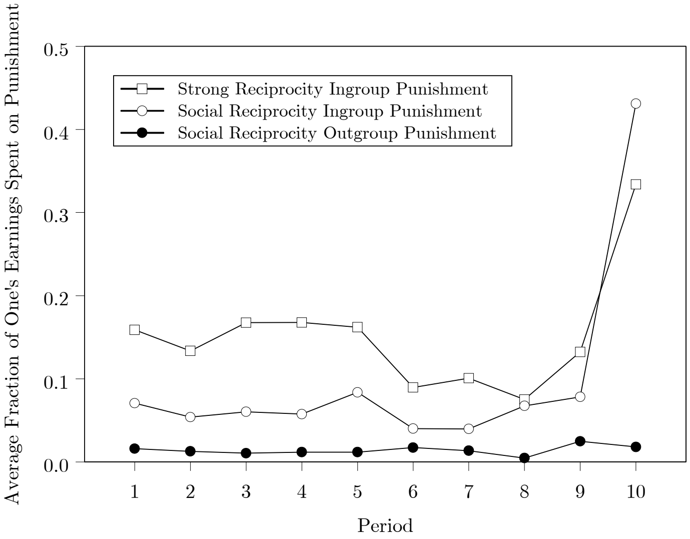


Figure 3: Average Expenditures on Punishment (VCM is the standard voluntary contribution mechanism, 4 groups; Strong Reciprocity is where only ingroup punishment is allowed, 6 groups; and Social Reciprocity is where players can punish both ingroup and outgroup, 14 groups).

KS=0.03, p=0.14.[5] Hence, we conclude that ingroup punishment is only marginally greater than outgroup punishment in our Social reciprocity treatment which begets the question: *Is there a common trigger of ingroup and outgroup punishment?* We return to this question in the next section.

However, to show that social reciprocity, as we define it, exists we simply need to show that outgroup punishment occurs, and it does. The simple test of whether the mean level of outgroup punishment including all the cases where people did not punish outgroup (but not controlling for contributions) is significantly greater than zero shows we can not reject the hypothesis that social reciprocity exists, t=8.57, p<0.01.

[5] The KS statistic refers to the Kolmogorov-Smirnov test for differences in cumulative distributions.

5 Punishment, Norm Violations, and Reciprocity

Now that we have established that social reciprocity occurs in our experiment, we wish to examine its origins. To do so we conducted a regression analysis of player punishment decisions. Because we hypothesize that the social reciprocity motivation for punishment is based on a simple normative impulse to punish rule breakers while strong reciprocal reasons to punish are based on the harm inflicted on another group member by a free rider, we model punishment choices as a two-step process that allows us to partially separate these two forces. Specifically, we hypothesize that social reciprocity is the reason that punishers get involved and, once involved, punishers justify how much punishment they inflict on targets by how much harm free riders inflict on the punisher. In the first step, the social reciprocity motive, contribution norm violators trigger whether a player gets involved or not. In the second step, the strong reciprocity motive, players who punish condition their punishment on the impact of the violation on their own welfare.

A natural way to model this decision process econometrically is to use the Heckman (1979) selection model. Based on the analysis of Fehr and Gächter (2000) who demonstrate that players who make contributions that are far from the group average (including those who contribute more than average) are more likely to be punished, we use one's absolute deviation from the group average contribution, $|Cont_{j,t} - Cont_{avg,t}|$, as our indicator of a broken commitment to a contribution norm. We test whether this deviation triggers socially reciprocal punishment. Conditional on punishing, we then measure the strong reciprocal aspect of punishment as the difference between the punisher's contribution and the target's contribution, $Cont_{i,t} - Cont_{j,t}$, which measures the impact that a free rider has on the payoff of the punisher.

Table 1: A Heckman Selection Model of Punishment

(dependent variable is $\boldsymbol{Punishment_{i,j,t}}$)

	Selection	**Punishment**
Dependent Variable	Punish or Not?	Punishment>0
Outgroup	-0.152	
	[-0.03]	
	(0.131)	
$\lvert Cont_{j,t}$ - $Cont_{avg,t}\rvert$	0.099	
	[0.02]	
	(0.012)**	
$\lvert Cont_{j,t}$ - $Cont_{avg,t}\rvert$ × *Outgroup*	-0.062	
	[-0.01]	
	(0.012)**	
Constant		6.696
		(4.425)
Outgroup		1.730
		(1.832)
$Cont_{i,t}$ - $Cont_{j,t}$		0.218
		$(0.130)^{+}$
($Cont_{i,t}$ - $Cont_{j,t}$ *)* × *Outgroup*		-0.204
		(0.176)
Rho		-0.303*
N	3920	410

Notes: (i) Regressions include time period fixed effects and cluster standard errors by group. (ii) (robust standard errors), [marginal effects for the selection probit]. (iii) $^{+}$ significant at the 10% level, * 5% level, ** 1% level.

Our experiment generates a panel of punishment choices (the dependent variable is $Punishment_{i,j,t}$: how much punisher i punishes target j in period t) which means that we should consider individual heterogeneity in our regressions. However, the standard procedure of including unobserved fixed or random effects for our participants is not the appropriate strategy in this situation because individual unobserved effects will be correlated with our selection indicator (see Wooldridge, 2002, Chp. 17). As an alternative strategy we cluster errors at the group level to account for the fact that group-member punishment decisions may not be independent and include time period fixed effects to control for idiosyncrasies in the progression of play (e.g., an endgame effect). Lastly, we consider only the punishment decisions of players in the social reciprocity treatment because of protocol differences between the standard punishment experiment, in which players monitor three other players, and the social reciprocity treatment where there are seven people to monitor, four of whom are in another group. The results of our analysis are summarized in Table 1.

In the first column of Table 1 appear the results of our first stage selection regression. As hypothesized, the social reciprocity decision to punish another player or not depends significantly ($p < 0.01$) on that player's deviation from the group average contribution. The more egregiously one breaks the group contribution norm, the more likely one is to be punished. Specifically, each deviation increases one's chances by 2% and, given the mean deviation is 3.95 EMUs, the average norm-breaker has an 8% chance of getting punished in the experiment. Also notice that, while there is no treatment effect of punishing outgroup (the coefficient on the *Outgroup* indicator is not significant), players react differently to norm violations that occur outside their groups. The interaction of our outgroup indicator variable and one's absolute deviation from the group average contribution indicates that people are approximately half as likely to punish deviations outside their groups. This result suggests that the norm violation motivation for punishment is weaker outside one's immediate group. However, this finding does not undermine the existence of social reciprocity, it simply suggests that social reciprocity conforms to standard notions of ingroup-outgroup behavior (Tajfel, 1981) in that ingroup violations are more salient.

Given one has decided to punish, the second column of Table 1 indicates that the ferocity of one's punishment depends significantly on the material harm the target imposes on the punisher. Each EMU that the punisher contributes above the target's contribution increases the target's punishment by 0.218 EMUs ($p < 0.10$). However, this motivation for punishment only appears to be strong within groups. As we might expect, strong reciprocity plays little role in the allocation of outgroup punishment. One can see this by first calculating the differential effect of contribution differences on outgroup punishing (i.e., the coefficient on the interaction), -0.204 and then summing this point estimate with the baseline effect, 0.218. The resulting figure, 0.014, which is very close to zero, is the effect of contribution differences between a monitor and a player in another group on the monitor's decision of how much to punish. The fact that this effect is essentially zero indicates that players rely on social reciprocity alone when punishing outside their groups. However, while this effect conforms to our hypotheses about punishment and reciprocity, it remains only suggestive because the coefficient on the interaction is not statistically significant at conventional levels ($p = 0.25$).[6]

6 The Efficiency of Social Reciprocity

We conjectured at the beginning of this paper that worlds in which social reciprocity existed would be more cooperative, in general, and would provide public goods more efficiently, in particular. In this section we illustrate that this conjecture is true and assess why social reciprocity facilitates collective action. Returning to Figure 2, we first note that contributions are significantly higher in the social reciprocity treatment confirming part of this conjecture – public goods are provided at higher levels when social reciprocity can be expressed. But our analysis so far does not allow us to claim they are provided more efficiently because we have not yet accounted for punishment expenditures and the costs of being punished.

We summarize the efficiency of providing the public good in Figure 4. In Figure 4 the vertical axis measures the ratio of the average net payoff for participants in a particular punishment treatment to the average payoff in the no-punishment control experiment. Hence, the heavy line at 1.0 is the benchmark efficiency of providing the public good when no punishment is allowed. In principle, punishment is socially worthwhile only if it generates efficiency gains over the situation in which no punishment is possible.

Early on, perhaps because players are becoming accustomed to the incentive structure, the efficiency of the two punishment treatments is lower than our benchmark, but the social reciprocity treatment is more efficient than the strong reciprocity treatment from the start. As the experiment progresses, the relative efficiency of both punishment regimes increase, but there is a noticeable difference in levels between the social reciprocity treatment and the strong

[6] Also notice that the correlation between the disturbances from our two stage regressions, rho, is significant indicating that selection is linked to allocation and our assumption about the structure of punishment is appropriate.

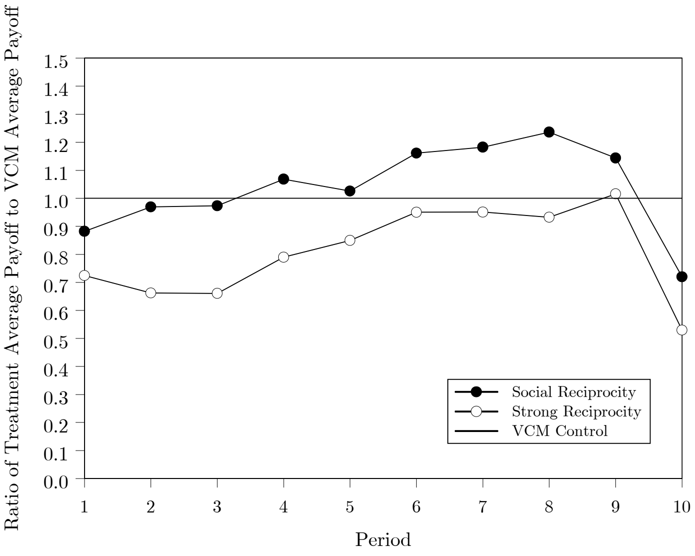


Figure 4: The Efficiency of Social Reciprocity (We graph the ratio of average payoffs in the treatments to the control. The divisor is the average payoff in the VCM, Strong Reciprocity is where only ingroup punishment is allowed, and Social Reciprocity is where players can punish both ingroup and outgroup).

reciprocity treatment. Payoffs are always substantially higher in the social reciprocity treatment than in the strong reciprocity treatment. Further, only in period nine is the strong reciprocity treatment briefly more efficient than the control, but starting in period four social reciprocity allows players to achieve sustained and growing efficiency gains over the control experiment. However, period ten is a disaster in both punishment conditions because free riders, without foresight, try to take advantage of the endgame and other players pummel them.

Why does social reciprocity increase the efficiency of public goods provision? We test two hypotheses that can explain the efficiency differences we see in Figure 4. First, if free riders are punished more severely in the social reciprocity treatment and punishment causes free riders to contribute more in the future, then contributions will be higher when outgroup punishment is allowed and social reciprocity is triggered. To test this hypothesis, we regress the total amount of punishment accruing to player i in round t on an indicator variable for the social reciprocity treatment, the player's deviation from the group average contribution (recall the discussion of Table 1), the player's own expenditure on punishment, and the interaction of the player's deviation from the group average contribution with our treatment indicator. If our first hypothesis about the efficiency gains is supported by our data, we expect a positive coefficient on the social reciprocity indicator.[7] As for our controls, we expect players with larger deviations from the average to accrue more punishment and we also expect that people who spend a lot on punishment themselves comply with the contribution norm and, therefore, are not punished much themselves.[8]

As we mentioned before, our experiment generates a panel of punishment decisions. Unlike the previous section, for this analysis we can account for individual heterogeneity with unobserved effects and we do so in equation (1) of Table 2. However, to be consistent with the methodology we use in the rest of our analysis, we also estimate this relationship using robust standard errors clustered by group and include time period fixed effects in equation (2). Lastly, because punishment can not be negative, we use the Tobit procedure to estimate both equations. The coefficient on the social reciprocity indicator is large, positive, and highly significant demonstrating that free riders are punished more severely when players monitor all potential norm violators. We also find that there is no significant correlation between the amount of punishment one receives and how much one spends to punish others, and, in accordance with our findings in Table 1, we see that larger deviations from the group average correlate significantly with more punishment, but this deviation matters less in the social reciprocity treatments. It is also interesting to see that the results are essentially identical when we cluster errors and use period fixed effects in equation (2). Taking stock, we see that free riders are punished more in the social reciprocity treatment, but we also need to show that free riders react prosocially to punishment.

[7] We might also be content with a positive coefficient on the interaction, but this would be highly unlikely given the results of Table 1 which showed that players worried less about deviations from the group average in other groups.

[8] See Carpenter et al. (2004) for an extensive discussion of the link between contributing and punishing.

Table 2: How Severely are Free Riders Punished?

(dependent variable is $\boldsymbol{Total\ Punishment\ Received_{i,t}}$)

	(1)	(2)
Social Reciprocity	18.388	18.769
	(4.384)**	(6.524)**
$Punishment\ Expenditure_{i,t}$	-0.159	-0.198
	(0.187)	(0.232)
$\lvert Cont_{j,t}$ - $Cont_{avg,t} \rvert$	3.569	3.485
	(0.548)**	(1.132)**
$\lvert Cont_{j,t}$ - $Cont_{avg,t} \rvert \times Social\ Reciprocity$	-1.855	-1.783
	(0.580)**	(0.655)**
Constant	-32.514	-31.724
	(4.358)**	(12.961)*
Includes individual random effects	Yes	No
Includes time period fixed effects and clusters errors by group	No	Yes
Wald chi^2	112	40
N	680	680

Notes: (i) Tobit regressions with lower bounds of zero. (ii) we report standard errors and robust standard errors in parentheses. (iii) $^+$ significant at the 10% level, * 5% level, ** 1% level.

Our contributions data is also a panel and, because we are interested in controlling for inertial effects when estimating the effect of punishment on contributions we include the lag of our dependent variable in the analysis. Of course this presents a problem because the lag will be correlated with an individual's unobserved effect (see Wooldridge, 2002, chp. 11). As above, the strategy we use to estimate the relationship between punishment and contributions is to incorporate time period fixed effects and cluster our errors at the group level. In equation (1) of Table 3 we report the results of regressing players' public contributions on the lags of their contributions and the punishment they received. We account for the fact that contributions are bound between 0 and 25 by using the Tobit estimator.

Equation (1) reveals two things about our pooled data. First, there is a lot of inertia in contributions. Second, overall, players respond to punishment by contributing significantly more in the future. This second fact confirms the assumption we made about the dynamics of contributing when we formulated our first explanation of why public goods are provided at higher levels in the social reciprocity treatment. Hence, we conclude that contributions are higher in the social reciprocity treatment because free riders who react prosocially to punishment are punished more. In fact, using equation (1) we can assess whether it "pays" to punish in the experiment. It costs 0.5 EMUs to inflict a 1 EMU punishment on a free rider. If this punishment occurs at the end of round one, the free rider will be expected to contribute 0.34 EMUs more in round two and the punisher's share of this increase is 0.17 EMUs (recall that $m = 0.5$). This does not seem like a very good deal. However, because of inertia, by the end of ten periods this unit of punishment will cause a 1.25 EMU total increase in contributions and the punisher's share of this total effect is 0.67 EMU - a much better deal.

A second hypothesis about why there are sustained efficiency gains in the social reciprocity treatment is that players respond more to punishment when more people are monitoring. That is, contributions might also be higher in the social reciprocity treatment because each unit of punishment has a greater effect in this treatment. In equation (2) of Table 3 we examine whether players react more to punishment in the social reciprocity treatment. The answer is yes. In fact, this regression indicates that increased punishment has no efficiency enhancing properties in the strong reciprocity treatment; all the benefits of punishment accrue to social reciprocity players.

To make our contributions regressions more consistent with our punishment regressions, in equation (3) we split people between those who contributed more than the group average last period and those who contributed less than the average last period (this leaves those who contributed at the average as the omitted category) and add the interactions with the treatment indicator. We continue to see that all the benefits of punishment seem to accrue to players in the social reciprocity treatment (which conforms to our hypothesis) but we also see that punishing those people who contributed more than the group average last period is very disruptive because this punishment causes these players to significantly reduce their contributions in the future.

Table 3: Do Free Riders Respond to Punishment?

(dependent variable is $\boldsymbol{Contribution_{i,t}}$)

	(1)	(2)	(3)
Social Reciprocity		2.038	0.213
		(3.519)	(1.433)
$Contribution_{i,t\text{-}1}$	0.743	0.844	0.872
	(0.145)**	(0.244)**	(0.138)**
Social Reciprocity × $Contribution_{i,t\text{-}1}$		-0.179	
		(0.292)	
$Punishment_{i,t\text{-}1}$	0.338	0.058	-0.050
	(0.147)**	(0.212)	(0.187)
Social Reciprocity × $Punishment_{i,t\text{-}1}$		0.324	0.449
		(0.274)*	(0.242)**
$Contribution_{i,t\text{-}1}\|Above\ Average$			-0.239
			(0.153)**
Social Reciprocity × $Contribution_{i,t\text{-}1}\|Above\ Average$			-0.030
			(0.191)
$Contribution_{i,t\text{-}1}\|Below\ Average$			-0.085
			(0.155)
Social Reciprocity × $Contribution_{i,t\text{-}1}\|Below\ Average$			-0.157
			(0.189)
Wald chi^2	114	458	890
N	720	720	720

Notes: (i) Tobit regressions include time period fixed effects and cluster standard errors by group. (ii) we report marginal effects not conditioned on being censored. (iii) (robust standard errors). (iv) $^+$ significant at the 10% level, * 5% level, ** 1% level.

We now summarize our efficiency results. Socially reciprocal worlds provide public goods more effectively *and* more efficiently. There are two reasons for this. First, because players will punish free riders outside their group, free riders are punished more severely in socially reciprocal worlds. Second, our players respond differently to punishment when social reciprocity is present. Specifically, increased punishment has much more of an effect on a free rider in the social reciprocity treatment. Perhaps, because they are punished more severely, social reciprocity players are quicker to learn that free riding is not acceptable.

7 Evidence Against Altruistic Punishment

So far we have spent our time differentiating social reciprocity from strong reciprocity and because we focus on outgroup punishment, tit-for-tat reasons for punishment have been controlled for in the design, but now we want to concentrate on showing that the results we call social reciprocity can not be explained by altruism either. We proceed by reviewing three pieces of evidence against altruism. The first bit of evidence is straightforward. Altruists would never punish in period ten because no benefits could follow for the other group members, yet there is substantial outgroup punishment in the last period (recall Figure 3).

While altruists would not punish outgroup on the last round, they may have a reason to punish in earlier periods. We have two additional pieces of evidence that suggest that the outgroup punishment that occurs in periods one through nine is mostly due to social reciprocity. First, if we can tie the behavior of those players who punish outgroup in period ten (social reciprocity for certain) to their behavior in periods one through nine then we can say something about who is most responsible for outgroup punishment during the rest of the game. We calculated the Spearman rank order correlation between how much a player punished outgroup in period ten and their propensity to punish outgroup in periods one through ten and found ρ=0.42 (p<0.01).[9] This correlation indicates that the players who punished in period ten were also the ones who had higher propensities to punish outgroup in the rest of the game. Hence, this suggests that most outgroup punishment comes from social reciprocators, not altruists.

Second, we conducted a post-experiment survey and asked specific questions about players motives to punish other players. In one question we asked:

> Which of the following sentences (if any) best describes your actions:
>
> a. I reduced the earnings of participants in the other group because I thought that in later rounds the earnings of participants in

[9] For each individual, regress one's punishment decisions on how much the outgroup target free rides. One's propensity to punish is the coefficient in this regression.

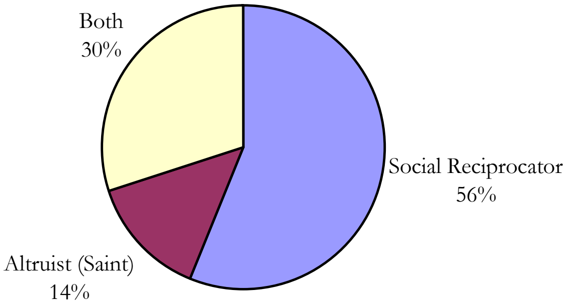


Figure 5: Stated Reasons for Outgroup Punishment (Social Reciprocators are people who said they punished outside their groups to get back at free riders, in general. Saintly Altruists are people who said they punished outgroup to help others. Those categorized as Both answered affirmatively to both responses).

the other group would be higher as a result.

b. I reduced the earnings of participants in the other group because I wanted to get back at those who did not contribute.

c. Both a. and b.

d. None of the above. Please explain:

The only reason players responded with (d) was because they did not punish anyone. Response (a) is the altruistic response and (b) is the social reciprocity response. The responses were distributed according to the pie chart in Figure 5. Social reciprocators outnumber altruists four to one and those who report being somewhat motivated by social reciprocity outnumber pure altruists approximately six to one.

We conclude that social reciprocity explains the majority of outgroup punishment. Tit-for-taters would never punish outside their groups, altruists would never punish in the last period, those social reciprocators who punish in the last period account for most of the outgroup punishment in the other nine periods, and simply asking people why they punish outgroup reveals that social reciprocity motivations outnumber altruistic motivations at least four to one. The existence of outgroup punishment and the efficiency gains to the community generated by social reciprocity leads to the following interesting result. Our data suggest that social reciprocity exists and is efficiency enhancing, but the efficiency gains are largely an unintended by-product because socially reciprocal agents do not necessarily punish with the purpose of increasing contributions in the future.

8 Towards A Model of Social Reciprocity

Our experimental results provide considerable support, in both the statistical and substantive senses of the word, for the proposition that free riders are punished both within *and* across groups. To provide some theoretical motivation for our results - we do not pretend, however, that no other rationale is possible or, in particular, that "prosocial preferences" cannot assume an important role in this context - we consider a "miniature social reciprocity game" (hereafter, MSR) consistent, in broad terms, with our experimental environment. Suppose that, at each moment in discrete time, "nature" chooses a "punishment network" of four individuals at random from a large (technically, infinite) population and then divides each foursome into pairs. MSR is then played in two stages. In the first, each of the two pairs plays its own public goods or voluntary contribution game, in which individuals must decide whether to contribute all or none of their endowment of 50 EMUs to a common pool with an MPCR of 50 percent. The normal form for each pair in the first stage is therefore:

	Contribute	Free Ride
Contribute	75, 75	37.5, 87.5
Free Ride	87.5, 37.5	50, 50

In the second stage, the choices of *all four* are then revealed to *all four*, after which *contributors* must decide (a) whether or not to enforce a "contribution norm" and punish free riders and, if so, (b) which free riders - ingroup, outgroup or both - to punish. We suppose, for purposes of simplification, that those who punish outsiders, the social reciprocators, cannot “pick and choose.” A contributor, for example, who is also committed to "norm enforcement" both within and across pairs and who is matched with three - one in and two out - free riders must sanction all three. Each punishment act is assumed to cost a contributor 10 EMUs, and to reduce a free rider’s payoff by 20 EMUs.

Consistent with the behavior that we observed in the lab, we further suppose that individuals in MSR are restricted to five pure strategies or behaviors: free ride and do not punish (F), contribute but do not punish (C), contribute and punish (just) ingroup free riders (I), contribute and punish (just) outgroup free riders (O) and contribute and punish both sorts of free riders (B).

We first note that MSR has two symmetric Nash equilibria or SNEs. The first, in which no one contributes and, therefore, no punishment is ever observed, is also MSR's unique subgame perfect equilibrium. In the second, however, the four participants randomize over the four contribution strategies, such that $p_I + 2p_O + 3p_B > 0.625$, where p_i is the likelihood that $i = F, C, I, O, B$ is played, and provides some support for the intuition that to deter free riding, the expected punishment costs must exceed some threshold. (For a derivation of this condition, see Appendix B.)

The second SNE is often dismissed, of course, because it fails the "backward induction test," the reason that punishment is often considered anomalous: if the punishment act is not costless, then no (implied) threat to sanction free riders should be credible, in which case there will be no reason, absent some sort of transformation of material outcomes into psychological ones, to contribute.

Punishment *is* observed, however, it cannot be rationalized as either conditional cooperation or "strong reciprocity" in the sense of Bowles and Gintis (2003). On the one hand, because the foursomes are dissolved at the end of each period, no individual is ever matched, absent a measure zero coincidence, with someone from a previous foursome of his or hers (In addition, under our experimental protocol, it was difficult, if not impossible, to tell who had punished whom, so it is not clear how much difference a possible rematch would make.) Such punishment cannot be understood, therefore, in terms of the Folk Theorem or the so-called "trigger strategies" that support conditional cooperation in some environments. On the other hand, the fact that at least some of this punishment is inflicted on outsiders implies that it cannot all be attributed to strong reciprocity, as Carpenter et al (2004) have underscored.

Within the framework of the model, then, the question of whether some, or even all, of the continuum of "all contribute" SNEs could meet some other, perhaps less restrictive, requirements for equilibrium becomes critical. In particular, we are interested in whether what we have called social reciprocity is, in a well-defined sense, *evolutionarily stable.* We should therefore first note that, as we have formalized it, MSR is an extension of what Axelrod (1984) first called the "Norms Game," a framework since featured in the research of Güth and Kliemt (1993), Binmore and Samuelson (1994) and Sethi (1996). In Sethi's (1996) reformalization, two players confront the usual prisoner's dilemma, after which each is free to punish the other, at some cost to him/herself, no matter what the other's first stage behavior. He then demonstrates that when each of the (eight) pure strategies available is identified with a sub-population of possible players, and a ninth sub-population, a set of best responders blessed with perfect recognition, is added, there will be two evolutionarily stable states (ESS) and one neutrally stable state (NSS). It is the first, monomorphic, ESS, in which "vengeful cooperators" comprise the entire population, that is most relevant here, not least because simulation exercises, based on the so-called replicator dynamic (Taylor and Jonker 1978), indicate that this outcome will be locally stable and that its basin of attraction could be substantial.

If Sethi's (1996) model provides a plausible explanation of the evolution of *strong* reciprocity in some environments, it remains to be seen whether *social* reciprocity can also sometimes survive selection pressures. Our approach here is not based on the ESS criterion - indeed, it isn't clear how ESS should be defined in this context (Broom, Cannings and Vickers 1997) - but rather the distinct notion of *drift compatible* population states (Binmore and Samuelson 1999). Our implementation is unusual, however, because we provide microfoundations for *both* the selection mechanism and drift function in terms of "learning" or "cultural transmission."

To this end, suppose that there are now five subpopulations associated with each of the five pure strategies in MSR and, in a convenient abuse of notation, denote their respective shares p_F, p_C p_I, p_O and p_B. To further streamline the exposition, we shall refer to their respective members as free riders, second order free riders, strong reciprocators, pure social reciprocators and social reciprocators. The evolution of population shares over time is then assumed to reflect two sorts of *reinforcement-based learning*, one more sophisticated and more common than the other. We suppose that sophisticated learners "sample and imitate" in the sense of Nachbar (1990), in which case the selection mechanism assumes the form of a scaled replicator dynamic, as confirmed below. The less sophisticated, on the other hand, are aspiration-driven as described in Carpenter and Matthews (2001), where the difference reflects how available information is, or is not used, either inside or outside the lab.

To be more precise, we suppose for the moment that time is marked in discrete intervals of length Δ and that at the end of each of these periods, a fraction $k\Delta$ of the entire population re-evaluates their present situations. A proportion $1-\theta$ of these, where θ is small, will sample another member of the population - that is, observe or somehow learn their behavior and outcome - and to switch or imitate whenever (a) the sampled payoff is higher and (b) the difference exceeds some switch cost c, the value of which is a random variable with uniform distribution over $[0,\bar{c}]$. To ensure the likelihood of a switch is always less than or equal to one, it is further assumed that $\bar{c} \geq 67.5$. A proportion θ, on the other hand, compare their current situation to some *aspiration level* a, the value of which is also a random variable, with uniform distribution over $[0,\bar{a}]$, where $\bar{a} \geq 87.5$. If one's payoff equals or exceeds this aspiration, the individual does not switch, but if it falls short, he or she "experiments" with another behavior. In the standard aspiration model (Binmore, Gale and Samuelson, 1995, for example), the probabilities that behaviors are adopted are assumed equal to their current population shares, but this implies that (a) these shares are observed and this information is processed and, more important, (b) the dissatisfied will sometimes "switch back" to the behavior that produced the dissatisfaction, neither of which seems desirable to us. Instead, we shall use a modified "no switch back dynamic" (Carpenter and Matthews 2001) here: individuals who have fallen short of their aspriations are assumed to switch to *another* pure strategy at random. It is this behavior that produces "drift" or "mutation" in our model.

Under these assumptions, the share p_i of the population committed to i evolves as follows:

$$\begin{aligned} p_i(t+\Delta) \quad = \quad & p_i(t) + (1-\theta)k\Delta\bar{c}^{-1}p_i[\sum_{j\neq i} p_j \max(0,\pi_i-\pi_j) - \\ & \sum_{j\neq i} p_j \max(0,\pi_j-\pi_i)] + \theta k\Delta\bar{a}^{-1}[0.25\sum_{j\neq i} p_j(\bar{a}-\pi_j) - p_i(\bar{a}-\pi_i)] \end{aligned} \tag{8.1}$$

The second term, for example, is the net increase in the share of i attributable to imitation. Of the $(1-\theta)k\Delta p_i$ percent of the population that is committed to i in period t who also reevaluate their performance, a fraction $p_j \max[0,\pi_j-\pi_i]$ will sample someone committed to $j \neq i$ whose outcome was better. Given the determination of switch costs, it then follows that a fraction $(1-\theta)k\Delta p_i\bar{c}^{-1}p_j \max[0,\pi_j-\pi_i]$ of the population will switch from i to $j \neq i$ as the result of imitation, and that the total number of "defections" will be $(1-\theta)k\Delta p_i\bar{c}^{-1}\sum_{j\neq i} p_j \max[0,\pi_j-\pi_i]$. In a similar vein, imitation will also cause a fraction $(1-\theta)k\Delta p_i \bar{c}^{-1}\sum_{j\neq i} p_j \max[0, \pi_i - \pi_j]$ of the population to switch *to* i.

The third term is the net increase in the share of sub-population i attributable to the less sophicated form of reinforcement: the likelihood that someone who is committed to $j \neq i$ falls short of his or her aspiration level is $(\bar{a}-\pi_j)/\bar{a}$, which implies that a fraction $\theta k \Delta \bar{a}^{-1}\sum_{j\neq i} p_j(\bar{a}-\pi_j)$ of the population will be dissatisfied with $j \neq i$, one quarter (0.25) of whom will then switch to i, and so on.

Since the bracketed expression in the second term collapses to the measure of "differential fitness" $\pi_i - \bar{\pi}$, where $\bar{\pi}$ is the average payoff for the population as a whole, (8.1) can be rewritten as:

$$\begin{aligned}\frac{p_i(t+\Delta)-p_i(t)}{\Delta} &= (1-\theta)k\bar{c}^{-1}p_i(\pi_i-\bar{\pi}) \\ &\quad +\theta k\bar{a}^{-1}[0.25\sum_{j\neq i} p_j(\bar{a}-\pi_i)-p_i(\bar{a}-\pi_i)]\end{aligned} \tag{8.2}$$

As $\Delta \to 0$, we have the continuous time version of (8.2):

$$\dot{p_i} = (1-\theta)\bar{c}^{-1}p_i(\pi_i-\bar{\pi})+\theta\bar{a}^{-1}[0.25\sum_{j\neq i} p_j(\bar{a}-\pi_i)-p_i(\bar{a}-\pi_i)] \tag{8.3}$$

after time has been rescaled (k alters the speed of population shares on their solution paths, but not the paths themselves).

In the special case where there is no drift ($\theta = 0$) or aspiration-driven "mutation," (8.3) is the standard replicator dynamic:

$$\dot{p_i} = \bar{c}^{-1}p_i(\pi_i-\bar{\pi}) \tag{8.4}$$

While our principal concern here is the behavior of (8.3), a brief discussion of the evolution of shares in the absence of drift provides some important intuition. We first note that the expected payoffs for the four subpopulations of contributors are a function of p_F, the proportion of first order free riders, alone:

$$\begin{aligned}\pi_C &= 75 - 37.5p_F \\ \pi_I &= 75 - 47.5p_F \\ \pi_O &= 75 - 57.5p_F \\ \pi_B &= 75 - 62.5p_F \end{aligned} \tag{8.5}$$

Since punishment is not costless, it comes as no surprise that for a fixed $p_F \neq 0$, those who punish more do worse: second order free riders, who do not punish, do better than strong reciprocators, who do not punish outside their group, and strong reciprocators do better than either sort of social reciprocator. What *is* unexpected is that the sometimes substantial differential between, for example, second order free riders and social reciprocators need not drive the latter to extinction. (This result does not turn, we should add, on the use of the replicator dynamic as a selection mechanism.) To understand this, we observe that the expected payoff for first order free riders or non-contributors is:

$$\pi_F = 27.5 + 22.5p_F + 60p_C + 40p_I + 20p_O \tag{8.6}$$

after substitution for $p_B = 1 - p_F - p_C - p_I - p_O$, which implies that first order free riders will, under some conditions, do worse than the social reciprocators. In this case, first oder free riders will sometimes be driven to extinction before social reciprocators and if this occurs, no contributor does better than the others, and the selection pressure on social reciprocators is eliminated.

Consider, for example, the situation in which the initial population is "balanced" - that is, $p_i(t = 0) = 0.20$ for all i. Second order free riders receive $75 - 37.5(0.20) = 67.5$ EMUs on average; strong reciprocators, 65.5; pure social reciprocators, 63.5; and social reciprocators, 62.5. First order free riders, on the other hand, receive just 56, which implies a mean population-wide payoff of 63. As the result of imitation, some first order free riders and social reciprocators would soon become second order free riders, a smaller number would instead become strong reciprocators, and a still smaller number would become pure social reciprocators. The first order free riders are more vulnerable, however - the likelihood that the payoff difference will exceed the switch cost is greater, in other words - in which case it is possible that their numbers will be driven to zero before those of the social reciprocators, which would eliminate the latter's fitness differential. Indeed, simulation of the RD from an initial balanced population reveals that, in rounded numbers, $p_F \Longrightarrow 0$, $p_C \Longrightarrow 0.34$, $p_I \Longrightarrow 0.26$, $p_O \Longrightarrow 0.22$ and $p_B \Longrightarrow 0.18$: that is, in the end, a little more than one third of the population will contribute but not punish, but $40(= 22 + 18)$ percent will be social reciprocators of one kind or another.

Two other properties of the evolution of population shares without drift also deserve mention. First, it should come as no surprise that, for some initial conditions, these shares will tend to an "all (first order) free rider" equilibrium in which $p_F \Longrightarrow 1$ and this is a desirable feature of the model: we do not always see cooperation and norm enforcement, either inside the lab or out. Second, the "all contribute" equilibrium is not unique: if the initial shares had been $p_F(0) = 0.10$, $p_C(0) = 0.15$, $p_I(0) = 0.20$, $p_O(0) = 0.25$ and $p_B(0) = 0.30$, for example, the population would evolve such that $p_F \Longrightarrow 0$, $p_C \Longrightarrow 0.18$, $p_I \Longrightarrow 0.23$, $p_O \Longrightarrow 0.27$ and $p_B \Longrightarrow 0.32$. It can be shown, in fact, that the relevant attractor is a subset of the shares that correspond to the component of mixed SNE in MSR.

There is reason to be concerned, however, that the all contribute equilibria of (8.3) are vulnerable to random drift. It should be noted, however, that while it isn't difficult to posit some "mutation" - the *massive* and *simultaneous* transformation of all kinds of contributors into first order free riders, for example - that would undo such equilibria, shocks of this sort are implausible. Rather, the issue here is whether or not the existence of small but persistent "noise" will push the population far from this component and toward the all free ride equilibrium. We are especially interested, for example, in whether outcomes in which all four contribute constitute a "hanging valley" (Binmore and Samuelson 1999) that is consistent with medium run equilibrium. In mechanical terms, our focus is on the behavior of (8.3) as θ tends to zero.

Closed form solutions to (8.3), expressed as a function of the drift paramater θ, are difficult (if not impossible) to obtain, however, so we report computed (with Maple) solutions for three values of θ, $0.01, 0.001$ and 0.0001, with the relevant eigenvalues, in Table 4, for the case in which $\bar{a} = \bar{c} = 100$.

	Noise Level			
	$\theta = 0.10$	$\theta = 0.01$	$\theta = 0.001$	$\theta = 0.0001$
p_F	0.044554	0.004632	0.000464	0.000046
p_C	0.295840	0.318817	0.321176	0.321411
p_I	0.247643	0.258326	0.259423	0.259532
p_O	0.212951	0.217130	0.217587	0.217633
p_B	0.199011	0.201095	0.201351	0.201378
Eigenvalues	-0.129493	-0.133335	-0.134637	-0.134777
	-0.020147	-0.001695	-0.000166	-0.000017
	-0.033345	-0.003000	-0.000296	-0.000030
	-0.028186	-0.002497	-0.000246	-0.000025
p_F	0.649904	0.976659	0.999772	0.999773
p_C	0.158916	0.010281	0.001003	0.001000
p_I	0.084318	0.005700	0.000557	0.000056
p_O	0.057382	0.003943	0.000386	0.000038
p_B	0.049478	0.003416	0.000334	0.000033
Eigenvalues	-0.043534	-0.121120	-0.124628	-0.124963
	-0.228658	-0.361696	-0.373634	-0.374864
	-0.106535	-0.312956	-0.323839	-0.224919
	-0.179235	-0.216488	-0.224182	-0.324885
p_F	0.308332	0.019156	0.001857	0.000185
p_C	0.337351	0.537293	0.551069	0.552410
p_I	0.159996	0.207271	0.209367	0.209565
p_O	0.104866	0.128402	0.129233	0.129311
p_B	0.089453	0.107878	0.108474	0.108529
Eigenvalues	0.024470	-0.037671	-0.034198	-0.033810
	-0.103800	0.005954	0.000654	0.000066
	-0.083164	-0.003308	-0.000320	-0.000032
	-0.029021	-0.005400	-0.000521	-0.000052

Table 4: Rest Points and Eigenvalues for MSR.

Table 4 reveals that under (8.3), MSR has *three* rest points, the properties of which seem robust with respect to the amount of drift. (We are confident, in other words, that the compositions of the population in the limit, as $\theta \Longrightarrow 0$, are close to these.) In the first, there are almost no free riders - in rounded numbers, the proportion is 0.4 percent when $\theta = 0.01$, and falls to 0.004 percent when $\theta =$

0.0001 - and the share of second order free riders, those who contribute but do not enforce norms, is about 32 percent in all three cases. Most important from the perspective of both our experimental results and model, however, almost 42 percent of the population are social reciprocators of one kind or another, and are therefore prepared to punish outsiders who do not contribute. This is, therefore, our "reciprocal equilibrium."

The second rest point corresponds to the backward induction equilibrium of MSR: the proportion of first order free riders runs from 97.7 percent when $\theta = 0.01$ to 99.9 percent when $\theta = 0.0001$, and no more than 0.7 percent of the population ever punish outsiders.

The third is similar to the first in the sense that there are almost no first order free riders, but there are also fewer social reciprocators - in each case, a little less than 24 percent - and more second order free riders. As Table 4 also reveals, however, this equilibrium is not stable: three of the four eigenvalues are negative, but the fourth is positive. The fact that is also small, however, has important implications, as seen below.

Figures 6 through 9 illustrate some possible solution paths. Figure 6, for example, plots the evolution of shares from a position of initial balance - that is, $p_i(0) = 0.20$ for all i - for the benchmark case $\theta = 0.01$, $\bar{a} = \bar{c} = 100$. As in the case of no drift, the population converges, rapidly, to the all contribute or reciprocity equilibrium. (In fact, the limit values are not far apart.)

What forces ensure that this outcome is stable, despite the continuous reintroduction of first order free riders to the population? It is useful to decompose the selective pressures that exist in this case. In the benchmark case, the normalized fitness differentials are:

$$\begin{aligned}
p_F(\pi_F - \bar{\pi}) &= 0.004632(61.408880 - 74.708783) = -0.000616 \\
p_C(\pi_C - \bar{\pi}) &= 0.318817(74.826300 - 74.708783) = +0.000375 \\
p_I(\pi_I - \bar{\pi}) &= 0.258326(74.779980 - 74.708783) = +0.000184 \\
p_O(\pi_O - \bar{\pi}) &= 0.217130(74.733660 - 74.708783) = +0.000054 \\
p_B(\pi_B - \bar{\pi}) &= 0.201095(74.710500 - 74.708783) = +0.000003
\end{aligned}$$

In the absence of mutation, then, the representative first order free rider does much worse than all four sorts of contributors, each of whom receives more than the population mean, so much so that despite the small size of their sub-

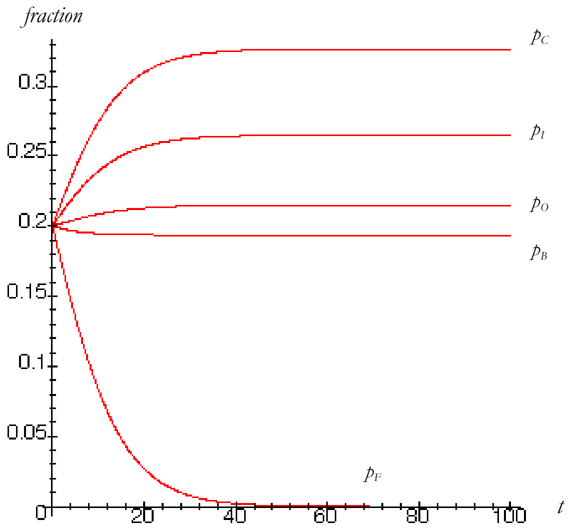


Figure 6: Evolution From an Initially Balanced Population.

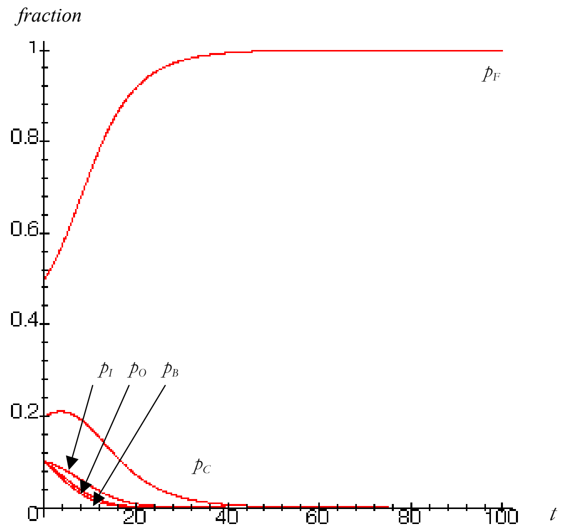


Figure 7: Almost Monotone Evolution to the No Contribution Equilibrium.

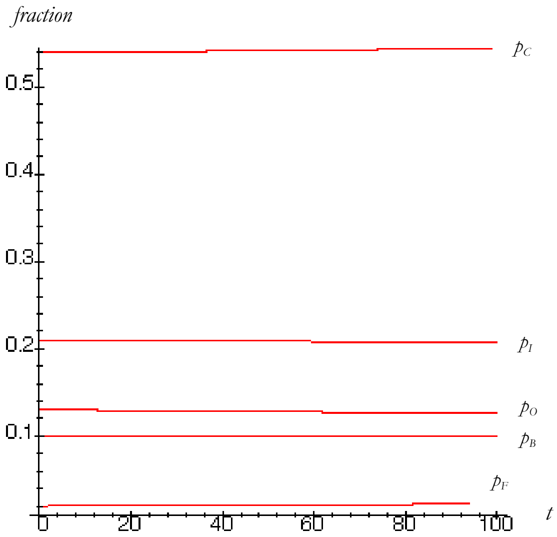


Figure 8: A Plateau Near the Unstable Equilibrium.

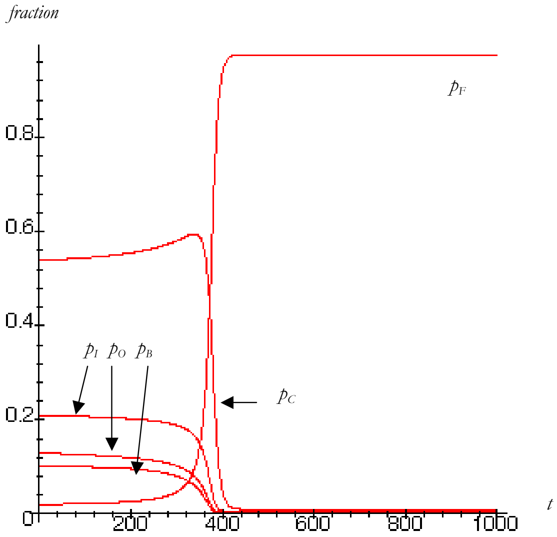


Figure 9: Falling Off a Plateau – the long run instability of the third equilibrium.

population, the decrease in their numbers is a substantial one. On the other hand, more than 60 percent of the free riders who switch as a result of imitation will become contributors who (also) do not punish and another 30 percent will become contributors who do not punish outsiders.

This in turn prompts the question: What prevents a population drift toward these two behaviors that would in turn favor free riders? The answer is found in the behavior of aspiration-based learners, which provides the required "offset." To see this, observe that the drift terms are:

$$\begin{aligned}
0.25\sum_{j\neq F} p_j(\bar{a}-\pi_j) - p_F(\bar{a}-\pi_F) &= 6.278116 - 0.178754 = +6.099361 \\
0.25\sum_{j\neq C} p_j(\bar{a}-\pi_j) - p_C(\bar{a}-\pi_C) &= 4.316353 - 8.025803 = -3.709450 \\
0.25\sum_{j\neq I} p_j(\bar{a}-\pi_j) - p_I(\bar{a}-\pi_I) &= 4.694057 - 6.514987 = -1.820929 \\
0.25\sum_{j\neq O} p_j(\bar{a}-\pi_j) - p_O(\bar{a}-\pi_O) &= 4.951284 - 5.486080 = -0.534796 \\
0.25\sum_{j\neq B} p_j(\bar{a}-\pi_j) - p_B(\bar{a}-\pi_B) &= 5.051406 - 7.760481 = -0.034186
\end{aligned}$$

As the numbers reveal, first order free riders are the one subpopulation to lose from imitation *and* to benefit from dissatisfaction. No less important, no contributors lose more unsophisticated learners than the second order free riders. To elaborate, while the likelihood (38.6% or 100-61.408880/100 $\approx$ 0.386) that the representative first order free rider falls short of his or her aspiration level exceeds that of the other four subpopulations, there are so few to start with that the absolute number of defections is small. On the other hand, the probabilities that a less sophisticated contributor will become disenchanted is smaller - from 25.2 percent for second order free riders to 25.3 percent for those who punish both insiders and outsiders - but because all four sorts, in particular second order free riders, are more numerous, the number of defections is also higher. Furthemore, because one quarter of all contributors who are dissatisfied will experiment with non-contribution, it is the first order free riders who benefit most. Second order free riders, on the other hand, are hurt most because more switch from, and few switch to, this behavior. Because the proportion of aspriation-based learners is just one percent, these cancel one another out.

In other words, the assumed nature of drift in this model implies that at the all contribution equilibrium, there is a constant flow of new first order free riders but because these non-contributors can expect to earn much less in an environment where almost all others contribute, and a substantial number of these are prepared to enforce contribution norms, there is also a constant, and equal, stream of defections.

Figure 7 depicts the evolution of population shares from the unbalanced initial condition in which first order free riders comprise half the population ($p_F = 0.50$), second order free riders another 20 percent ($p_C = 0.20$), and strong, pure social and social reciprocators 10 percent each ($p_I = p_O = p_B = 0.10$). In this case, there is rapid and almost monotone convergence to the no contribution equilibrium.

Figures 8 and 9, on the other hand, illustrate one of the more "exotic" possibilities that follow from the introduction of drift. The initial point is chosen close to the third, unstable, equilibrium, $p_F = 0.02$, $p_C = 0.54$, $p_I = 0.21$, $p_O = 0.13$ and $p_B = 0.10$, and Figure 8 plots the evolution of population shares *over the same time horizon* as Figures 6 and 7, a period of time more than sufficient to "settle down" in those cases. It seems that there is an almost imperceptible drift in the population, *from* first order free riders *toward* second order free riders, and perhaps a plateau of sorts. Figure 9, which provides a much longer run perspective on the same dynamics, demonstrates that this conclusion would be premature: in short order, the share of first order free riders explodes, while the share of second order free riders, which exceed 50 percent, collapses and, in the end, a stable no contribution equilibrium is established. In this case, the model exhibits what is in effect a régime shift, from a situation in which almost all contribute to one in which almost no one does. While we did observe a collapse of this sort in one or two experimental sessions, a "rebirth" of the contribution norm also followed.

Given a fixed value of θ, each of the stable rest points is hyperbolic, so that small changes in the values of either $\bar{a}$ or $\bar{c}$ will have small changes on equilibrium shares, but it is important to ask what would happen if, for example, one of the parameters doubled in size. The issue is moot, of course, in the absence of drift, since aspiration levels are (in this case, at least) irrelevant and the switch cost affects the speed of evolution but not its path. To this end, Tables 5 and 6 present some comparative statics for the model's two stable equilibria.

		Switching Cost	
	$\bar{c} = 100$	$\bar{c} = 150$	$\bar{c} = 200$
p_F	0.004632	0.006909	0.009159
p_C	0.318817	0.317525	0.316247
p_I	0.258326	0.257726	0.257133
p_O	0.217130	0.216882	0.216638
p_B	0.201095	0.200958	0.200824
p_F	0.976659	0.964687	0.952496
p_C	0.010281	0.015567	0.020958
p_I	0.005700	0.008621	0.011593
p_O	0.003943	0.005961	0.008013
p_B	0.003416	0.005164	0.006941

Table 5: The Comparative Statices of Switching Costs.

		Aspiration Upper Bound	
	$\bar{a} = 100$	$\bar{a} = 150$	$\bar{a} = 200$
p_F	0.004632	0.009172	0.011406
p_C	0.318817	0.316239	0.314967
p_I	0.258326	0.257129	0.256538
p_O	0.217130	0.216636	0.216395
p_B	0.201095	0.200824	0.200693
p_F	0.976659	0.968470	0.964291
p_C	0.010281	0.013895	0.015741
p_I	0.005700	0.007698	0.008717
p_O	0.003943	0.005323	0.006028
p_B	0.003416	0.004612	0.005222

Table 6: The Comparative Statices of Dissatisfaction.

The results show that when there is not much drift, the equilibrium shares are not much affected, even when the sizes of $\bar{a}$ and $\bar{c}$ double, from 100 to 200. Furthermore, the effects on the equilibrium shares are consistent with intuition. An increase in the value of $\bar{c}$, for example, increases the amount of "inertia": to induce the less successful to switch, the difference in outcomes must be more substantial. This in turn reduces the selective pressure on less successful behaviors, which implies that their equilibrium shares will decrease, and this is indeed what happens. In the reciprocity equilibrium, the proportions of all four sorts of contributors become smaller - the differences, however, are from the third decimal place onward - while the proportion of first order free riders increases, from 0.46 percent to 0.91. For the same reason, the share of first order free riders in the no contribution equilibrium falls 2.5 percent, to 95.2 percent, while the shares of all four sorts of contributors increase a little bit.

In a similar vein, an increase in $\bar{a}$ increases the likelihood that an individual will fall short of his or her aspriation no matter how successful (in relative terms, at least) their MSR outcomes, so that here, too, one would expect the shares of "favored subpopulations" to decrease, and vice versa, and the results in Table 6 confirm this.

To be consistent with our experimental data, however, it must also be the case that contributors survive under more than some small and perhaps contrived set of initial conditions. That is, the first equilibrium should be stable and have a substantial basin of attraction. Given the dimension of (8.3), a pictorial characterization is difficult, and some sort of compression is needed. To this end, Figure 10 plots the evolution of first and second order free riders for the set of initial conditions $p_F(0) = 0.1, 0.2..., 1$ and $p_i(0) = (1 - p_F(0))/4$ for all $i \neq F$ - that is, for the case where the initial shares of the four sorts of contributors are equal. It shows that when the initial share of first order free riders is less than 25 percent or so, reinforcement-based "evolution" will drive the population to the all contribute equilibrium, but when the initial share exceeds this, the population instead moves toward the no contribution equilibrium. In some cases, the process is slow and exhibits the same sudden shifts illustrated in Figure 9: on the path labelled A, for example, there is a sudden turnaround in the fortunes of first order free riders at time t_1.

Figure 11 illustrates the evolution of the same two population shares under a different set of initial conditions, $p_C(0) = 0.1, 0.2..., 1$ and $p_i(0) = (1 - p_C(0))/4$ for all $i \neq C$, a condition that equalizes the numbers of free riders and each of the three sorts of contributors who punish. In this case, when the initial share of second order free riders is a third or less, first order free riders almost vanish, consistent with the intuition that for non-contributors to flourish, the combined shares of those prepared to punish such behavior must be smaller than some threshold value. Otherwise, there is sometimes slow and roundabout evolution toward the no contribution equilibrium.

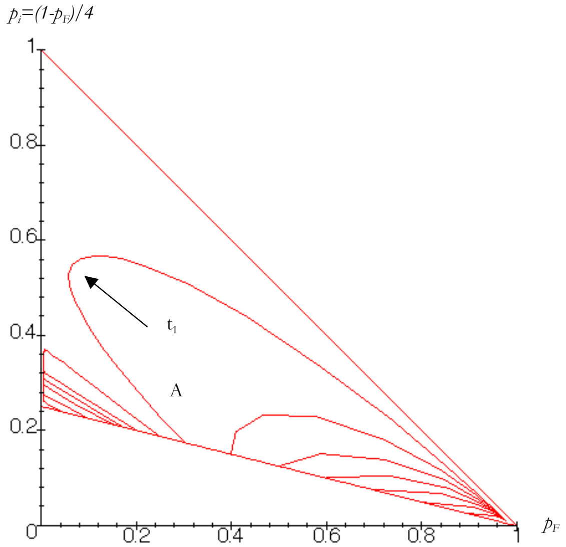


Figure 10: A View of the Basins of Attraction (p_F is the fraction of free riders and p_i is the equal share of all the other strategies).

The crucial common feature of Figures 10 and 11 is that the survival of reciprocity, both strong and social, is not unusual or limited to a small neighborhood of the all contribute equilibrium.

9 Conclusion

> Who sees not that vengeance, from the force alone of passion, may be so eagerly pursued as to make knowingly neglect every consideration of ease interest and safety? David Hume, *An Enquiry Concerning the Principles of Morals*, 1751

This paper provides an integrated - experimental and theoretical - perspective on "social reciprocity," which we define as the willingness to enforce norms with little regard to group affiliation or social distance. Furthermore, it shows that such behavior should be distinguished from more familiar (conditional, strong) forms of reciprocity, and also from altruism. In some sense, then, the model rationalizes the now familiar claim that "it (sometimes) takes a village"

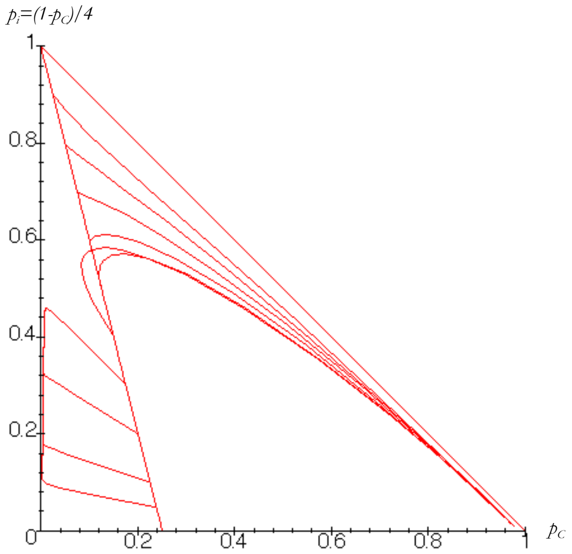


Figure 11: Another View of the Basins of Attraction (p_C is the fraction of unconditional cooperators and p_i is the equal share of all the other strategies).

but also, on the basis of the second stable equilibrium, the observation that even villages will sometimes fall short of the mark.

We do not pretend, of course, that ours is a complete characterization, and at least three possible extensions come to mind. First, at a conceptual level, the paper considers negative but not positive manifestations of reciprocal behavior but there are some environments in which the latter are more important. This in turn underscores the need to consider more specific "frames" or situations. How, for example, does social reciprocity matter in the workplace?

Second, at a theoretical level, the model is intended to serve as a point of departure, and not a canonical treatment. The two sorts of learners in the model, for example, are described as sophisticated and unsophisticated, but even the former's rule is a simple one, and it remains to be seen whether our results extend to models with other, perhaps more elaborate, rules. It is possible, for example, that under other rules, the evolution of shares would be consistent with both the sudden collapse of contribution norms, as in our model, but also with their rebirth, as we observed in some experimental sessions. The role of social preferences or, for that matter, the "ordered beliefs" that characterize psychological games, also require exploration, not least their influence on the determination of initial conditions.

Third, in terms of the experiment, it remains to be seen whether similar results obtain with other subject pools - workers, for example – for which reciprocal behavior is more important.

10 References

Acheson, J. (1988): The Lobster Gangs of Maine. Hanover: University Press of New England.

— (1993): "Capturing the Commons: Legal and Illegal Strategies," in The Political Economy of Customs and Culture: Informal Solutions to the Commons Problem, ed. by T. Anderson, and R. Simmons. Lanham: Rowman & Littlefield, 69-84.

Andreoni, J. (1988): "Why Free Ride? Strategies and Learning in Public Good Experiments," Journal of Public Economics, 37, 291-304.

Axelrod, R. (1984): "An Evolutionary Approach to Norms," American Political Science Review, 80, 1095-1111.

Binmore, K., J. Gale, and L. Samuelson (1995): "Learning to Be Imperfect: The Ultimatum Game," Games and Economic Behavior, 8, 56-90.

Binmore, K., and L. Samuelson (1994): "An Economist's Perspective on the Evolution of Norms," Journal of Institutional and Theoretical Economics, 150, 45-63.

— (1999): "Evolutionary Drift and Equilibrium Selection," Review of Economic Studies, 66, 363-393.

Borofsky, G., G. Stollak, and L. Messe (1971): "Sex Differences in Bystander Reactions to Physical Assault," Journal of Experimental Social Psychology, 7, 313-318.

Bowles, S., J. Carpenter, and H. Gintis (2001): "Mutual Monitoring in Teams: Theory and Evidence on the Importance of Residual Claimancy and Reciprocity," mimeo.

Bowles, S., and H. Gintis (2003): "The Evolution of Strong Reciprocity," Theoretical Population Biology, forthcoming.

Broom, M., C. Cannings, and G. Vickers (1997): "Multi-Player Matrix Games," Bulletin of Mathematical Biology, 59, 931-952.

Carpenter, J. (2002): "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and Collective Action."

Carpenter, J., and P. Matthews (2001): "No Switchbacks: Rethinking Aspiration-Based Dynamics in the Miniature Ultimatum Game."

Carpenter, J., P. Matthews, and O. Ong'ong'a (2004): "Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms," Journal of Evolutionary Economics, forthcoming.

Craig, B., and J. Pencavel (1995): "Participation and Productivity: A Comparison of Worker Cooperatives and Conventional Firms in the Plywood Industry," Brookings Papers: Microeconomics, 121-160.

Fehr, E., and U. Fischbacher (2001): "Reputation and Retaliation."

— (2003): "The Nature of Human Altruism," Nature, 425, 785-791.

— (2004): "Third Party Punishment and Social Norms," Evolution and Human Behavior, forthcoming.

Fehr, E., and S. Gaechter (2000): "Cooperation and Punishment in Public Goods Experiments," American Economic Review, 90, 980-994.

Ghemawat, P. (1995): "Competitive Advantage and Internal Organization: Nucor Revisited," Journal of Economics and Management Strategy, 3, 685-717.

Gintis, H. (2000): "Strong Reciprocity and Human Sociality," Journal of Theoretical Biology, 206, 169-179.

Gueth, W., and H. Kliemt (1993): "Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation, and Moral Attitudes," Metroeconomica, 45, 155-187.

Heckman, J. (1979): "Sample Selection Bias as a Specification Error," Econometrica, 47, 153-161.

Isaac, R. M., J. Walker, and S. Thomas (1984): "Divergent Evidence on Free-Riding: An Experimental Examination of Possible Explanations," Public Choice, 43, 113-49.

Kandel, E., and E. Lazear (1992): "Peer Pressure and Partnerships," Journal of Political Economy, 100, 801-17.

Latane, B., and J. Darley (1970): The Unresponsive Bystander: Why Doesn't He Help? New York: Appleton-Century-Crofts.

Ledyard, J. (1995): "Public Goods: A Survey of Experimental Research," in The Handbook of Experimental Economics, ed. by J. Kagel, and A. Roth. Princeton: Princeton University Press, 111-94.

Moir, R. (1998): "Spies and Swords: Costly Monitoring and Sanctioning in a Common-Pool Resource Environment," mimeo.

Nachbar, J. H. (1990): "Evolutionary Selection Dynamics in Games: Convergence and Limit Properties," International Journal of Game Theory, 19, 59-89.

Olson, M. (1965): The Logic of Collective Action. Cambridge: Harvard University Press.

Ostrom, E. (1992): Crafting Institutions for Self-Governing Irrigation Systems. San Francisco: ICS Press.

Ostrom, E., R. Gardner, and J. Walker (1994): Rules, Games and Common-Pool Resources. Ann Arbor: University of Michigan Press.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants with and without a Sword: Self-Governance Is Possible," American Political Science Review, 86, 404-17.

Page, T., and L. Putterman (2000): "Cheap Talk and Punishment in Voluntary Contribution Experiments," mimeo.

Sampson, R., S. Raudenbush, and F. Earls (1997): "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy," Science, 277, 918-924.

Sefton, M., R. Shupp, and J. Walker (2000): "The Effect of Rewards and Sanctions in Provision of Public Goods," mimeo.

Sethi, R. (1996): "Evolutionary Stability and Social Norms," Journal of Economic Behavior and Organization, 29, 113-40.

Shotland, L., and M. Straw (1976): "Bystander Response to an Assault: When a Man Attacks a Woman," Journal of Personality and Social Psychology, 34, 990-999.

Tajfel, H. (1981): Human Groups and Social Categories. Cambridge: Cambridge University Press.

Taylor, P., and L. Jonker (1978): "Evolutionary Stable Strategies and Game Dynamics," Mathematical Biosciences, 40, 145-56.

Wooldridge, J. (2002): Econometric Analysis of Cross Section and Panel Data. Cambridge: The MIT Press.

11 Appendix A: Experiment Participant Instructions

You have been asked to participate in an experiment. For participating today and being on time you have been paid $5. You may earn an additional amount of money depending on your decisions in the experiment. This money will be paid to you, in cash, at the end of the experiment. When you click the BEGIN button you will be asked for some personal information. After everyone enters this information we will start the instructions for the experiment.

During the experiment we will speak in terms of Experimental Monetary Units (EMUs) instead of Dollars. Your payoffs will be calculated in terms of EMUs and then translated at the end of the experiment into dollars at the following rate: 30 EMUs = 1 Dollar.

In addition to the $5 show-up fee, each participant receives a lump sum payment of 15 EMUs at the beginning of the experiment.

The experiment is divided into 10 different periods. In each period 8 participants are divided into two groups of 4. The composition of the groups will remain the same for the entire experiment. Therefore, in each period your group will consist of the same four participants.

Each period of the experiment has three stages.

Stage One

At the beginning of every period each participant receives a 25 EMU endowment. In Stage One each of you will decide how much of the 25 EMUs to contribute to a group project and how much you want to keep for yourself. You are asked to contribute whole EMU amounts (i.e. a contribution of 5 EMUs is alright, but 3.85 should be rounded up to 4). Your payoff and the payoff of everyone else in your group will be determined by how much each member contributes to the group project and how much each member keeps.

To record your decision, you will type EMU amounts in two text-input boxes, one for the group project labeled GROUP ALLOCATION and one for yourself labeled PRIVATE ALLOCATION. These boxes will be yellow. Once you have made your decision, there will be a green SUBMIT button that will record your decision.

After all the participants have made their decisions, each of you will be informed of your gross earnings for the period.

GROSS EARNINGS

Your Gross Earnings will consist of two parts:

(1) Earnings from your Private Allocation. You are the only beneficiary of EMUs you keep. More specifically, each EMU you keep increases your earnings by one.

(2) Earnings from the Group Project. Each member of the group gets the same payoff from the group project regardless of how much he or she contributed. The payoff from the group project is calculated by multiplying 0.5 times the total EMUs contributed by the members of your group.

Your Gross Earnings can be summarized as follows:
1 × (EMUs you keep) + 0.5 × (Total EMUs contributed by your group)

Let's discuss three examples.

Example 1: Say each member of your group contributes 15 of their 25 EMUs. In this case, the group total contribution to the project is 4 × 15 = 60 EMUs. Each group member earns 0.5 × 60 = 30 EMUs from the project. The gross earnings of each member will then be the number of EMUs kept, 25-15 = 10, plus the earnings from the group project, 30 EMUs, for each member. Hence, each member would earn 10+30=40 EMUs.

Example 2: Now say everyone in the group contributes 5 EMUs. Here the group total contribution will be 20 and each member will earn 0.5 × 20 = 10 EMUs from the group project. This means that the total earnings of each member of the group will be 20 (the number of EMUs kept) plus 10 (earnings from the group project) which equals 30 EMUs.

Example 3: Finally, say three group members contribute all their EMUs and one contributes none. In this case, the group total contribution to the project is 3 × 25 = 75 EMUs. Each group member earns 0.5 × 75 = 37.5 EMUs from the project. The three members who contributed everything will earn 0+37.5 = 37.5 EMUs and the one member who contributed nothing will earn 25+37.5 = 62.5 EMUs.

Stage Two

In stage two you will be shown the allocation decisions made by all the other participants, and they will see your decision. Also at this stage you will be able to reduce the earnings of other participants, if you want to, and the other participants will be able to reduce your earnings. You will be shown how much each member of your group kept and how much they allocated to the group project. You will also be shown how much each member of the other group kept and how much they contributed to their group project. Your allocation decision will also appear on the screen and will be labeled YOU.

Please remember that the composition of your group remains the same during each period and therefore every person in your group during this period will also be in your group next period.

At this point you will decide how much (if at all) you wish to reduce the earnings of the other participants. You reduce someone's earnings by typing the number of EMUs you wish to spend to reduce that person's earnings into the input-text box that appears below that participant's allocation decision.

For each EMU you spend you will reduce the earnings of the other participant by 2 EMUs. You can spend as much of your accumulated earnings as you wish to reduce the earnings of the other participants.

Consider this example: suppose you spend 2 EMUs to reduce the earnings of a participant in the other group, you spend 9 EMUs reducing the earnings of a participant in your group, and you don't spend anything to reduce the earnings of the remaining participants. Your total cost of reductions will be (2+9+0) or 11 EMUs. When you have finished you will click the blue DONE button.

How much a participant's gross earnings are reduced is determined by the total amount spent by all the other participants in the session. If a total of 3 EMUs is, then this person's earnings will be reduced by 6 EMUs. If the other participants spend 4 EMUs in total, the person's earnings would be reduced by 8 EMUs, and so on.

Stage Three

In stage three, you will be shown the total EMUs spent on reductions by each other participant. You will then be able to spend an additional amount of money to reduce the earnings of the other participants, if you choose to do so.

Again, for each EMU you spend you will reduce the earnings of the other participant by 2 EMUs. You can spend as much of your accumulated earnings as you wish to reduce the earnings of each of the other participants. When you have click the blue DONE button.

Nobody's earnings will be reduced below zero by the other participants. For example, if your gross earnings were 40 EMUs and the other participants spent 50 EMUs to reduce your earnings, your gross earnings would be reduced to zero and not minus sixty.

Your NET EARNINGS after the third stage will be calculated as follows:

(Gross Earnings from Stage One) - (2 × the number of EMU spent on reductions directed towards you) - (your expenditure on reductions directed at other participants)

If you have any questions please raise your hand. Otherwise, click the red FINISHED button when you are done reading.

12 Appendix B: MSR's Symmetric Nash Equilibria (SNE)

We shall first show that the two common profiles identified in the text are indeed SNEs, and then show that no others are possible. The argument that the first profile - that is, the case in which all four choose to free ride - satisfies this criterion is trivial, so we shall focus on the second, in which all four randomize over the four pure contribution strategies. Consider the common mixture $\sigma^i = (0, p_C, p_I, p_O p_B)$ for all $i = 1, ..., 4$. There is no incentive for j to deviate to some other mixture over the four contribution strategies - she would continue to earn 75 - so that attention can be limited to strategies of the form $\sigma^j = (p_F^j, p_C^j, p_I^j, p_O^j, p_B^j)$ where $p_F^j > 0$, with payoff $\pi^j(\sigma^j, \sigma^i, \sigma^i, \sigma^i)$. It follows that $\pi^j = p_F^j \pi_F^j + (1 - p_F^j)75 = 75 + p_F^j(\pi_F^j - 75)$, where π_F^j is what j can expect to earn as a unilateral free rider, and therefore that there will be no incentive to deviate from σ^i if $\pi^j < 75$ or, substituting in the previous expression, $\pi_F^j < 75$. Under what circumstances will this condition be met? That is, under what conditions can the unilateral free rider expect to receive less than 75? We first observe that she will earn 87.5 with likelihood $p_C(p_C + p_I)^2 + p_O(p_C + p_I)^2 = (p_C + p_O)(p_C + p_I)^2$, where the first term is the product of the likelihood p_C that her partner will choose to contribute but not punish and the likelihood that both members of the outgroup will either contribute but not punish or contribute and punish insiders. Following similar logic, she will receive 67.5 with likelihood $2p_C(p_C + p_I)(p_O + p_B) + p_I(p_C + p_I)^2 + 2p_O(p_C + p_I)(p_O + p_B) + p_B(p_C + p_I)^2$, 47.5 with likelihood $p_C(p_O + p_B)^2 + 2p_I(p_C + p_I)(p_O + p_B) + p_O(p_O + p_B)^2 + 2p_B(p_C + p_I)(p_O + p_B)$, and 27.5 with likelihood $p_I(p_O + p_B)^2 + p_B(p_O + p_B)^2$. Gathering terms, we have:

$$\begin{aligned}
\pi_F^j \quad = \quad & 87.5p_C(p_C + p_I)^2 + 87.5p_O(p_C + p_I)^2 + 135p_C(p_C + p_I)(p_O + p_B) \\
& +67.5p_I(p_C + p_I)^2 + 135p_O(p_C + p_I)(p_O + p_B) + 67.5p_B(p_C + p_I)^2 \\
& +47.5p_C(p_O + p_B)^2 + 95p_I(p_C + p_I)(p_O + p_B) + 47.5p_O(p_O + p_B)^2 \\
& +95p_B(p_C + p_I)(p_O + p_B) + 27.5p_I(p_O + p_B)^2 + 27.5p_B(p_O + p_B)^2
\end{aligned}$$

or, after factoring:

$$\begin{aligned}
\pi_F^j &= (p_C + p_O)[87.5(p_C + p_I)^2 + 135(p_C + p_I)(p_O + p_B) + 47.5(p_O + p_B)^2] \\
&\quad (p_I + p_B)[67.5(p_C + p_I)^2 + 95(p_C + p_I)(p_O + p_B)27.5(p_O + p_B)^2] \\
&= (p_C + p_O)[87.5(p_C + p_I) + 47.5(p_O + p_B)][p_C + p_I + p_O + p_B] \\
&\quad (p_I + p_B)[67.5(p_C + p_I) + 27.5(p_O + p_B)][p_C + p_I + p_O + p_B]
\end{aligned}$$

Since $p_C + p_I + p_O + p_B = 1$, this can be rewritten:

$$\begin{aligned}
\pi_F^j &= (p_C + p_O)[87.5(p_C + p_I) + 67.5(p_I + p_B)] \\
&\quad +(p_O + p_B)[47.5(p_C + p_O) + 27.5(p_I + p_B)] \\
&= 87.5(p_C + p_O) + 67.5(p_I + p_B) - 40(p_O + p_B) \\
&= 87.5p_C + 67.5p_I + 47.5p_O + 27.5p_B
\end{aligned}$$

It follows, therefore, that $\pi_F^j < 75$ if and only if:

$$87.5p_C + 67.5p_I + 47.5p_O + 27.5p_B < 75$$

or, since $p_C = 1 - p_I - p_O - p_B$ in this case:

$$20p_I + 40p_O + 60p_B > 12.5$$

or:

$$p_I + 2p_O + 3p_B > 0.625$$

which is the condition in the text.

The remaining candidates for SNE are those in which players randomize over free riding and one or more of the contribution strategies. To show that none of these are in fact viable, we note that attention can first be restricted to strategies of the form $\sigma^i = (p_F, 1-p_F, 0, 0, 0)$: if there is some positive likelihood that each of the others will free ride, then profiles that sometimes call for the punishment of free riders will fare worse than those that do not. The members of this restricted set can also be ruled out, however, since in the absence of punishment, contribution is dominated.