“Social Reciprocity”

by

Jeffrey Paul Carpenter and Peter Hans Matthews

July, 2002

MIDDLEBURY COLLEGE ECONOMICS DISCUSSION PAPER NO. 02-29

DEPARTMENT OF ECONOMICS
MIDDLEBURY COLLEGE
MIDDLEBURY, VERMONT 05753

http://www.middlebury.edu/~econ

SOCIAL RECIPROCITY

Jeffrey Paul Carpenter

Department of Economics
Middlebury College
Middlebury, Vermont 05753

jpc@middlebury.edu

Peter Hans Matthews

Department of Economics
Middlebury College
Middlebury, Vermont 05753

peter.h.matthews@middlebury.edu

Current Version: June 2002
Comments Welcome

Abstract: We conduct a survey and find that 47% of respondents state they would sanction free riders in a team production scenario even though the respondent was not personally affected and no direct benefits could be expected to follow an intervention. To understand this phenomenon, we define social reciprocity as the act of demonstrating ones disapproval, at some personal cost, for the violation of a widely-held norm (for example, don't free ride). Social reciprocity differs from reciprocity because social reciprocators punish all norm violators, regardless of group affiliation or whether or not the punisher bears the costs. Social reciprocity also differs from altruism because, while the latter is an outcome-oriented act benefiting someone else, the former is a triggered response not conditioned on future outcomes. To test the robustness of our survey results, we run a public goods experiment that allows players to punish each other. The experiment confirms the existence of social reciprocity and additionally demonstrates that more socially efficient outcomes arise when reciprocity can be expressed socially. Further we find that most subjects who punish do so to discipline transgressors and helping others is largely a positive externality. Finally, to provide some theoretical foundations for social reciprocity, we show that generalized punishment norms survive in one of the two stable equilibria of an evolutionary public goods game with selection drift.

JEL Classification Numbers: C79, C91, C92, D64, H41

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

SOCIAL RECIPROCITY*

> 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

1. Introduction

Despite strong incentives to free ride on others efforts, individuals in groups facing social dilemmas (e.g., public goods, common pool resources, and team production) appear to be particularly adept at attenuating incentive problems without external intervention. Communities often develop rules that make contributing and free-riding transparent (Ostrom [1992]), but perhaps more importantly community members are also often willing to incur costs to monitor and sanction behavior that benefits the individual but harms the group. 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. But the punishment of free riders can also emerge in situations where the individual costs clearly outweigh any possible benefits if (1) group members are motivated by reciprocity, in which case they reciprocate the costs imposed by free-riders, or if (2) group members are motivated by altruism, in which case they punish to benefit others.

However, it is unclear where the boundaries of retribution, reciprocity, and altruism end, especially when communities grow and group boundaries blur. For this reason we propose the following taxonomy of the reciprocal and altruistic motivations for punishment. To clarify concepts we borrow the framework for characterizing normative behavior developed in Elster [1989]. On one hand, norm-driven behavior can be separated from other types of behavior because norms do not cause people to act for instrumental reasons. That is, norms are not outcome-oriented. In other words, people who follow norms react to stimuli without considering the possible strategic implications of their actions. For example, norm-driven people who intervene when a stranger needs help do not do so because they anticipate some reward will accrue to them or anyone else (i.e., they don't consider possible future benefits) but because it seems like the right thing to do. On the other hand, outcome-oriented people operate via backwards induction and perpetrate acts now

* We thank Okomboli Ong'ong'a for research assistance and the members of the junior faculty research seminar at Middlebury College for thoughtful comments on an earlier version of this work. Middlebury College and the National Science Foundation (SES-CAREER 0092953) provided financial support.

with the hope that these acts will have predictable and beneficial future consequences.

Limiting our attention to social sanctions and punishment, we identify five distinct behavioral types by differentiating reciprocal from altruistic reasons for punishment, and norm-driven from outcome-oriented motivations for action. We think of the now classic *Tit-for-Tat-er* or conditionally cooperative person as being reciprocal and outcome-oriented. The reason is that models of conditional cooperation usaully rely on games being repeated and agents having sufficiently long time horizons so that tit-for-tat is established, via the Folk Theorems, as enlightened self-interest.

Likewise, we realize that many altruistic acts occur for instrumental reasons. Specifically, *Samaritans* punish asocial individuals because they anticipate that punishment will cause the latter to change their behavior in the future and this will in turn benefit others. We differentiate Samaritans from what *Saints* by contrasting the formers' altruistic, but instrumental, use of punishment from the latters' norm-driven or moral reasons to punish. Saints punish asocial types because asocial people inflict harm on society, in general, rather than because saints want to help any person in particular. In short, the saintly norm says, "Help people in distress even if doing so entails punishing some people."

This leaves norm-driven, reciprocal motivations for punishing free-riders. We argue that *Reciprocators* follow a norm that requires subscribers to punish deviations from widely held conventions. That is, reciprocators do not punish with the future consequences of their actions in mind, but to sanction rule-breakers. The following table summarizes our taxonomy:

		Reasons to Punish	
		Reciprocity	Altruism
Motivations for Action	Outcome-Oriented	Tit-for-Tat-ers	Samaritans
	Norm-Driven	Reciprocators Strong Reciprocity Social Reciprocity	Saints

In this paper we are interested in understanding the origins, limits, and social implications of individuals who incur costs to express their disapproval of asocial behavior. However, rather than surveying all the explanations of punishment defined above, we wish to focus on building a case for norm-driven reciprocal explanations. Specifically, we hope to shed light on why individuals engage in costly acts to punish asocial agents when the punisher has been directly harmed and when neither the punisher nor her group has been directly harmed.

To be as precise as possible, we further distinguish between two types of norm-driven reciprocal behavior based on group boundaries. *Strong Reciprocators* (Bowles and Gintis [1999], 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 without regard to group boundaries and may therefore punish free riders in group 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 or the social distance between punisher and rule breaker. Further, we hypothesize that the trigger for punishment by strong reciprocators is the cost implicitly imposed by a free-rider on the group (e.g., a public good is provided at a lower level), while 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.

Another way of thinking about the relationship between strong reciprocity and social reciprocity is that social reciprocity is a generalization of strong reciprocity when group boundaries are fuzzy. There are plenty of examples of fuzzy group boundaries. One example would be a neighborhood within a city. In many cases, it is not obvious where one neighborhood starts and another ends. Another example may arise in team production when multiple teams occupy the same shop floor. Strong reciprocity dictates that members of a specific team punish free riders from that team only and disregard what happens in other teams. By contrast, social reciprocity requires people to sanction all shirkers regardless of what team they work in.

More examples will help illustrate what we call social reciprocity. Psychologists and sociologists have long been interested in understanding personally costly acts that benefit others. A subset of this vast literature, motivated by the infamous 1964 murder of Kitty Genovese in Queens, focuses on bystander intervention in situations in which someone is breaking an obvious rule. Two studies (Borofsky et al. [1971], Shotland and Straw [1976]) demonstrate 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 Borofsky et al., pooling across treatments, 29% of bystanders intervene in a situation in which two confederates of the experimenters stage an altercation that escalates in to a physical fight. Shotland and Straw show that 65% of bystanders intervene when a stranger (a male confederate) assaults a woman (female confederate). However, they also show that only 19% intervene when the two confederates appear to be married. The explanation they offer is in terms of costs to intervene. Bystanders reason that strangers are much more likely to run off when confronted than are husbands who are more likely to stay and fight.

A third example of social reciprocity comes from Latane and Darley [1970]. In this experiment, subjects are asked to wait in a room to be interviewed. Also in the room is a confederate who, when the experimenter leaves, steals what remains of the show-up fee fund. The dependent variable is the probability that subjects report the theft when the experimenter returns. Notice, at this point the subjects have been paid their show-up fee and therefore suffer no loss from the theft which means strong reciprocity should not be triggered, nor should they expect a reward for turning the confederate in and so instrumental reasons for acting should not play a role. Furthermore, the potential cost of turning in the confederate is high (i.e. the confederate may retaliate). Therefore, neither tit-for-tat reasoning nor strong reciprocity can explain why someone would report the theft. Regardless, in 50% of the cases that 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 more efficiently, 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 members will be caught more often and punished more widely than societies characterized by no reciprocity or societies based on standard, tit-for-tat or strong notions of reciprocity.

While the psychological experiments mentioned above provide evidence that people will intervene in potentially costly situations in which people break rules, we are more interested in situations that have direct economic importance. To see whether people will intervene in a production environment such as the team production example mentioned above, we conducted a survey with college students which placed the respondent in a team production setting. We asked students to respond to questions about how to deal with free riding in three different vignettes. In one scenario, *Strong Reciprocity*, the respondent worked in a team of four where one other team member shirked and this imposed a cost on the team. In a second scenario, *Social Reciprocity (Low Cost)*, the respondent, again, worked in a team, but this time was asked whether he or she would intervene and sanction a shirking member of a different team whos shirking had no effect on the respondents compensation. Lastly, the third scenario was identical to the second except we added a line that stated that the shirker intended to retaliate if turned in. We call this the *Social Reciprocity (High Cost)* scenario.[1]

Table 1 highlights the important results from this survey. We find that there is consensus across the treatments that the shirker should be punished even in the

[1] See Appendix A for the exact wording of each vignette.

two situations in which the respondent would have no material interest in seeing the shirker punished. Further, in the strong reciprocity scenario we see that most people (90would intervene and sanction the shirker themselves. As one would expect if the population is heterogeneous and some people consider the benefits of punishing, the fraction of people who say they would intervene in either social reciprocity scenario is substantially lower. Regardless, we find that a substantial number of people say they would intervene in an economically important situation in which they were neither harmed nor would they expect future benefits.

[Insert Table 1 About Here]

As a general test of our hypothesis that intervention is caused by norm violations, we analyzed other responses of those who said they would intervene in either of the social reciprocity scenarios. Specifically, we also asked whether the respondent agreed that by shirking, the person broke an unstated rule (i.e., a norm or convention) and whether the respondent would personally feel badly if he or she did not intervene (i.e., is the intervention norm internalized). The general results for our sample of 49 students are summarized in the following ordered probit regressions - the data were coded using a seven-point lIkert scale - with standard errors reported in parantheses.

$$\begin{aligned} Punish\ (Low\ Cost) = {} & 0.58BreakNorm - 0.06Internalized \\ & (0.21) \qquad\qquad\quad (0.21) \\ & PseudoR^2 = 0.06 \\ & \chi^2 = 10.12 \end{aligned}$$

$$\begin{aligned} Punish\ (High\ Cost) = {} & 0.60BreakNorm + 0.16Internalized \\ & (0.18) \qquad\qquad\quad (0.22) \\ & PseudoR^2 = 0.10 \\ & \chi^2 = 15.36 \end{aligned}$$

In line with the social reciprocity hypothesis, respondents say they will punish shirkers in the both cost scenarios because the shirkers are violating a work norm ($p < 0.01$), but it does not appear that the social reciprocity norm has been internalized by the respondents. Feeling badly for not intervening does not cause our average respondent to punish someone shirking in another work team.

While we feel our survey results are suggestive, the potential difference between punishing when called on to do so and simply saying one would punish raises concerns about hypothetical biases. To bolster our 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. The following five sections 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 eighth 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. Section 9 then offers some concluding thoughts.

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 extract 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 Gaechter [2000] who confirm a 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 Gaechter showed that free riders are punished and contributions increase.

The work of Fehr and Gaechter 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 test 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 [2001] 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, together with our survey results from section 1, suggest 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. tit-for-tat, Samaritanism, and sainthood). 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 sanction 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 are not outcome-oriented and who dont 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 (Strong) in which players could monitor and sanction other members of their group (6 groups), and our social reciprocity treatment (Social) in which players could monitor and punish all the other players in a session, but they only benefited from their groups contribution to a public good (14 groups).

The payoff function for the social reciprocity treatment was similar to the peer 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 the other group. Likewise, *ingroup punishment* occurs when members of a group punish each other. In the VCM treatment no sanctions were allowed. In the Strong treatment no outgroup sanctions were allowed and players saw only the contributions of their group members. But, in the Social 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

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

players 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 $\sigma^k \equiv (\sum \sigma_i^k)/n$ 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 so that the contribution rate also measured how efficiently the public good was provided.

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 incredibleand therefore can not 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 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 is no possible future benefit.

In the Social 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 players decision. After all players were done distributing sanctions, the experiment moved

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

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.

[Insert Figure 1 About Here]

4. Does Social Reciprocity Exist?

The first question we wish to answer 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 sanctioned another player at least once. Specifically, 82% of our subjects sanctioned ingroup and 50Hence, a preliminary look at our data suggests half our participants exhibit some degree of social reciprocity - clearly a significant number.

Figure 2 presents a summary of our three treatments. The vertical axis measures (1) the fraction of the individual endowment (25 EMUs) contributed to the public good, on average, and (2) the fraction of a punishing players gross earnings spent on sanctioning other players, 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] in particular). 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 using means tests and Kolmogorov-Smirnov tests for distributional differences in the pooled data.[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 Social treatment react less to the endgame. Despite the end-game effect, our first major result is that *social reciprocity exists and leads to increased individual contributions to a public good.*

[Insert Figure 2 About Here]

Concerning punishment expenditures, the first thing to notice is that our Strong treatment seems to elicit more ingroup punishment than the Social treatment. However, one should be careful drawing this conclusion because, as was just mentioned, contributions are significantly higher in the Social treatment. This means less pun-

[4] In fact, the smallest t value was 3.95 testing differences in the Social and Strong means ($p = 0.0001$) and the lowest KS statistic was 0.16 ($p = 0.0005$) which compares the distributions of Social and Strong contributions.

ishment was warranted.[5] Our second observation about punishment is, within the Social treatment, it appears players spend more resources punishing ingroup than outgroup players. However, while this appears to be the case when looking at figure 2, 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, $KS = 0.03, p = 0.14$. Hence, we conclude that ingroup punishment is only marginally greater than outgroup punishment in our Social treatment which begets the question: *Is there a common determinant of ingroup and outgroup punishment*? We return to this question below.

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, $t = 8.57, p < 0.01$.

We also have other evidence pointing to the existence of social reciprocity. In both punishment treatments, players spend a significant portion of their earnings to punish free riders (both in- and outgroup) in period ten when there clearly can be no effect on future contributions. This fact should eliminate and outcome-oriented reasons for punishment including Samaritanism.

[5] Without controlling for contributions, ingroup punishment was significantly higher in the Strong treatment, $t = 3.52, p = 0.0004; KS = 0.19, p < 0.01$.

5. Propensities to Punish

Now that we have established that social reciprocity exists, we wish to examine its origins. In particular, we are interested in what triggers punishment, and in getting a more accurate measure of the population distribution of social reciprocators. To do so we conducted a regression analysis of player punishment decisions. Our experiment generates a panel of punishment choices and to account for individual heterogeneity we add random effects to all relevant regressions.

The first question we ask is, *what are the determinants of punishment*? The answers to this question are summarized in the first equation in table 2. $FreeRide$ measures the fraction of the endowment a player keeps, Outgroup equals one when the player is deciding how much to punish some outgroup player, and Strong is one to differentiate players in the Strong treatment where outgroup punishment was not allowed. Hence, our reference decision maker is a player deciding how much to punish someone else in his or her group when participating in the Social treatment.

[Insert Table 2 About Here]

The coefficient on $FreeRide$ is highly significant ($p < 0.0001$) indicating punishers take into account whether, and how much someone else free rides. Specifically, a player can expect that if they contribute one EMU less to the public good ($\Delta FreeRide = 0.04$) the average punisher will spend 0.07 EMUs to punish which will reduce the players gross payoff by 0.14 EMUs. The differential gross payoff of free riding by one more EMU is 0.50 EMUs (the difference between keeping an EMU and contributing it) and including punishment costs the net payoff is $0.50 - 0.14 Pun_{avg}$, where Pun_{avg} is the average number of punishers per free rider. If everyone else in the free riders group punishes, free riding still pays, but if enough outgroup players punish too, free riding wont pay. The average number of players that punish free riders in the Strong treatment is 1.08 and in the Social treatment, on average, 1.80 players punish free riders.[6] Doing the calculations we see that in either treatment the expected incremental return on free riding is positive, but it is less in the Social treatment.

There are two other things to notice about this regression. First, the Outgroup coefficient is highly significant ($p < 0.01$) and negative. This suggests that, controlling for how much people free ride, players punish significantly less outside their group. Second, the coefficient on Strong is positive but not significant ($p > 0.26$) which implies punishment is doled out the same in our two ingroup treatments after controlling for how badly a player free rides. This second result is important because it indicates that people think about punishing outside their group in addition to punishing in their group, not instead of. The significant negative outgroup coefficient is important because it suggests that players care less about free riders

[6] For now we define a free rider as a player who contributes a third or less of her EMU endowment.

who do not have a direct material impact on them. This presents a puzzle, *why do socially reciprocating players punish less outside their group*?

This is the point at which our story about the determinants of punishment becomes more complicated. As our results show, punishment seems to be additively separable into ingroup and outgroup punishment. The question is whether the causes of the two components are the same or different. Clearly, strong reciprocity cant explain outgroup punishment because players in the other group inflict no harm on outgroup punishers, but can social reciprocity explain both. We now turn to an analysis of this question. If the answer is affirmative, then we can conclude that all punishment is triggered by breaking the rules. If the answer is negative, then we must conclude that the two components of punishment have different causes. For example, ingroup punishment may be triggered by the harm caused by free riders and outgroup punishment may be triggered by the violation of a contribution norm. We begin this portion of our analysis be confirming differences in ingroup and outgroup punishment.

Call a player's *propensity to punish* free riders the marginal effect of free riding on a players decision to punish. To measure player's propensities to punish we calculated individual punishment coefficients from the following regression for each of our three punishment conditions:

$$PunishmentExpenditure_{ij} = \alpha + \beta_i FreeRide_{ij} + \epsilon$$

Figure 3 summarizes our results.[7] In figure 3 we report the cumulative distributions of our individual punishment propensity measures for the Strong treatment, Social ingroup punishment, and Social outgroup punishment. Fifty-four percent of the propensities are significantly different from zero at the ten percent level in the Strong treatment, 58% are significant for the Social ingroup punishment decisions, and 25% are significant for the Social outgroup decisions.

[Insert Figure 3 About Here]

It appears that the propensities to punish ingroup members in the Social treatment are similar to the propensities to punish in the Strong treatment. In fact, there are more non-negative propensities in the Social treatment (74% compared to 67%), but the average propensities are not statistically different ($z = 0.1, p > 0.92$) and, perhaps more importantly, the distributions are not different ($KS = 0.14, p > 0.82$). Overall, most players differentially punish free riders in their group more, the more they free ride.

It is also interesting to compare the propensities to punish in ones group to the propensity to punish outside ones group. Half our participants decided to

[7] Two outliers have been deleted to create figure 3. One β in the Social ingroup condition equalled 171.48 and a second in the Strong condition equalled 27.82. However, they are included in the analysis.

not punish outside their group compared to eighteen percent who did not punish inside their group. The difference in the average propensity to punish is significant ($z = 4.16, p < 0.01$) and the ingroup propensities are distributed significantly to the right of the outgroup propensities ($KS = 0.39, p < 0.001$). We conclude that individual players, as well as the average player, punish less severely outside the group than inside the group.

This analysis also provides us with a second, more conservative measure of social reciprocity - thirty-eight percent of our participants had positive propensities to punish outgroup free riders.[8]

Our measures of the propensities to punish confirm and add to the mystery of why social reciprocators react less to outgroup free riding. We now hope to shed some light on this puzzle. Recall, the hypothesis we are now working from is that strong reciprocity motivates players to punish free riders because of the harm they inflict and social reciprocity *additionally* motivates players to punish free riders simply because they break the rules. To analyze this theory we added a dummy variable to our punishment analysis. $FRdummy$ takes the value one when a player contributes less than one third of his or her endowment to the public good. Our interpretation is that $FRdummy$ indicates that someone has violated the contribution norm while $FreeRide$ measures the material harm done by free riding. If players react more to the violation of the norm than to the material harm done, the FR dummy should be just as good a predictor of punishment as is $FreeRide$.

Table 2 presents the results. Equation one is the same regression we reported above. In equation two we substitute the $FRdummy$ variable for $FreeRide$ to test whether one variable predicts punishment better than the other. The results are mixed, the coefficient on $FRdummy$ is less, but it is no less significant than $FreeRide$. Further, the magnitudes of the other regressors and R^2s dont radically change suggesting norm violation explains punishment as well as the material harm caused by free riding.

[Insert Table 2 About Here]

[8] Another way to assess differences in punishment behavior across treatments is by comparing what we call predispositions to punish free riders. A players *predisposition to punish* free riders is how much they tend to punish free riders after controlling for how badly the free rider breaks the norm. To calculate players' predispositions to punish we add individual fixed effects to the standard punishment regression. The relative distribution of players predispositions is similar to the distributions of propensities. The Strong and Social ingroup predispositions are essentially the same ($z = 0.68, p > 0.49; KS = 0.20, p > 0.43$) while the Social ingroup predispositions are distributed higher than in the Social outgroup condition ($z = 2.04, p < 0.5; KSs = 0.45, p < 0.001$) indicating players are more watchful of their group members than they are of outsiders.

To have a closer look, we separate the data and analyze ingroup punishment and outgroup punishment individually. [9] Regressions three and four confirm our suspicions about the two components of reciprocity. Only the material harm inflicted by free riding ($FreeRide$) explains punishment inside a group. However, both the material harm done by free riding and the fact that a rule has been broken trigger outgroup punishment. In fact, breaking the norm is the more serious reason for outgroup punishment (i.e. the coefficient on $FRdummy$ is more highly significant). Also notice that there are two ways to interpret the $FreeRider$ variable. So far we have interpreted $FreeRider$ as measuring the cost imposed by a free rider on reciproctors, but a second interpretation is that it just measures how egregiously someone violates the norm.[10] Given both $FreeRider$ and $FRdummy$ are significant causes of outgroup punishment, the most reasonable interpretation of $FreeRide$ might be the second because we have no reason to believe the cost inflicted on the other group motivates outgroup punishment, but perhaps how badly one breaks the norm does matter to social reciprocators.

The following are our conclusions about punishment and reciprocity. First and foremost, social reciprocity exists. Our conservative measure states that 38% of people go out of their way to punish norm infractions caused by people who have no material impact on them, and our least conservative measure notes that 50% of people punish outside their group at least once. Second, we find that social reciprocity may explain both ingroup and outgroup punishment if we allow that social reciprocators are mostly motivated by the violation of a norm, but they also react to how egregiously free riders violate the norm.

[9] Regressions one and two include the data from the Strong treatment while regression three is only on the Social ingroup data and regression four is only on the Social outgroup data.

[10] In fact, Mudd [1968] and Mudd [1972] show that, instead of reciprocating costs imposed, punishment depends on (i.e. is a linear function of) the degree to which the perpetrator violates an accepted norm.

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 assess to what extent this conjecture is true and why it might be true. Returning to figure 2, we first note that contributions are significantly higher in the Social 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.

[Insert Figure 4 About Here]

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 treatment is more efficient than the Strong treatment from the start. As the experiment progresses, both punishment regimes increase relative efficiency, but there is a noticeable levels difference between Social and Strong. Payoffs are always substantially higher in the Social treatment than in the Strong treatment. Only in period nine is the Strong 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 social reciprocators pummel them.

Why does social reciprocity increase the efficiency of public goods provision? We offer two explanations for why efficiency is noticeably higher in the Social treatment. First, free riders are punished more severely in the Social treatment than in the Strong treatment. As mentioned above, on average, more players punish a free rider in the Social treatment (1.80 versus 1.08) and the total punishment received by free riders in the Social treatment is higher. Table 3 lists, by period, the average payoff reduction charged to free riders in the two treatments. We consider two definitions of free riding. Based on previous peer monitoring experiments which find that contributing less than the average triggers punishment, the first definition we use is a free rider is someone who contributes less than the group average. The second definition is more arbitrary. Here we say free riders are people who contribute less than one-third of their endowment.

[Insert Table 3 About Here]

Beginning with the arbitrary measure, we see that punishment is relatively stable, except for the last two periods, and that free riders are punished more severely in the Social treatment in seven of the ten periods. Using the deviations from the average standard, free riders are still punished more severely in the Social treatment (six of ten periods), but the difference is not as extreme as when we considered the arbitrary measure. We also conducted Wilcoxon and Kolmogorov-Smirnov tests on the pooled punishment data to confirm that free riders are punished at higher levels when social reciprocity is possible.

A second reason why there are sustained efficiency gains in the Social treatment is that players respond more to punishment when more people are watching. In table 4 we report the results of regressing players public contributions on lags of their contributions and punishment received.[11] Equation one 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. Equation two answers any questions remaining about figure 2. Controlling for the amount of punishment received, players are more cooperative in the social reciprocity treatment. This alone provides a reason why social reciprocity leads to efficiency gains. In equation three we examine why players are more cooperative in the Social treatment. The answer is that an increase in punishment in the Social treatment has a much bigger effect when compared with the Strong treatment. In fact, equation three says increased punishment has no efficiency enhancing properties in the Strong treatment, all the benefits of punishment accrue to Social players.

We now summarize our efficiency results. Figure 4 illustrates that our conjecture about the efficiency of social reciprocity is confirmed - 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 treatment than in the Strong treatment. Perhaps, because they are punished more severely, Social players are quicker to learn that free riding is not acceptable.

[Insert Figure 4 About Here]

[11] Because contributions are constrained from below by zero and from above by the endowment, 25 EMUs, we use the tobit procedure in addition to adding individual random effects.

7. Evidence Against Altruistic Punishment

So far we have spent much of 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 we now 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. First, Samaritans 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.

So far we established that Samaritans dont punish outgroup in the last round, but they may have a reason to punish in earlier periods. We have two additional pieces of evidence that suggest while there is an altruistic incentive to punish in periods one through nine, most of the punishment we see is 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 $\rho = 0.42$ ($p < 0.01$). 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 Samaritans.

However, we also need to rule out saintly punishment. To this end, 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 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 saintly altruistic response and (b) is the social reciprocity response. The responses were distributed according to the pie chart in figure 5. Social reciprocators outnumber saints four to one and those who report being somewhat motivated by social reciprocity outnumber pure saints approximately six to one.

[Insert Figure 5 About Here]

We conclude that social reciprocity explains the majority of outgroup punishment. Tit-for-taters would never punish outside their groups, Samaritans 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 saintly 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 byproduct because socially reciprocal agents dont punish with the purpose of increasing contributions in the future.

8. Towards A Model of Social Reciprocity

Our experimental results are unambiguous: in both the statistical and substantive senses of the word, there is *significant* punishment of free riders, both within *and* across groups. To provide some theoretical foundations for these results, we construct a "miniature social reciprocity game" or MSR consistent in broad terms with our experimental environment. Suppose that at each moment in continuous time, nature selects four individuals at random from a large (technically, infinite) population and then divides each foursome into pairs. In the first stage of MSR, each of the two pairs plays its own public goods 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	$87.5, 37.5$	$50, 50$

The choices of all four individuals are then revealed to all and, in the second stage, contributors must decide (a) whether or not to enforce a "contribution norm" and punish free riders, and (b) which free riders - ingroup, outgroup or both - to punish. We shall assume, for purposes of simplification, that those who punish outsiders, the social reciprocators, cannot "pick and choose": for example, a contributor who is committed to norm enforcement both within and across groups and is matched with three - one ingroup and two outgroup - 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.

We shall also suppose that individuals are restricted to five pure strategies: free ride and do not punish, contribute and do not punish, contribute and punish ingroup free riders, contribute and punish outgroup free riders, and contribute and punish all free riders. That is, free riders cannot or will not punish contributors or other free riders and contributors cannot or will not punish other contributors, restrictions that we believe are consistent with the motivations of our experimental subjects. Consider, for a moment, MSR's two symmteric Nash equilibria or SNEs. In the first, no one contributes, but in the second, all four randomize over the four contribution strategies such that $p_3+2p_4+3p_5 > 0.625$, consistent with the intuition that to deter free riders, the expected punishment must exceed some lower bound. (For a derivation of the SNEs, see Appendix C.)

The first equilibrium is observationally equivalent to MSR's unique subgame perfect equilibrium and, to the extent that the latter constitutes a benchmark of sorts, the reason that punishment of either kind is considered anomalous: if the punishment act is not costless, then no threat to sanction free riders in the second stage should be considered credible, in which case there should be no reason to contribute in the first.

Punishment *is* observed, however, and it is important to recall that it has two important properties. First, it cannot be rationalized on the basis of the Folk Theorem(s): the foursomes are dissolved at the end of each period, so that no one should expect to be matched with a particular individual from a previous foursome or, for that matter, with a particular member of the population as a whole, in subsequent rounds. In other words, there are no opportunities to engage in "conditionally reciprocal" behavior. Second, as Carpenter and Matthews [2002] have underscored, if some of this punishment is inflicted on outsiders, it cannot all be attributed to strong reciprocity (Gintis [2000], Bowles and Gintis [2000]), the predisposition to cooperate and punish those who have reduced one's own welfare.

From the perspective of the model, then, the question is then whether some elements of the continuum of SNEs could satisfy some other, perhaps less restrictive, notion of equilibrium. To be more precise, we are interested in whether socially reciprocal behavior is ever *evolutionarily* stable. It is therefore important to note that, as we have formalized it, MSR is an extension of what Axelrod [1984] first called the "norms game," a framework also featured in the work of Güth and Kliemt [1993], Binmore and Samuelson [1994] and Sethi [1996]. In Sethi's [1996] two person (re)formalization of Axelrod [1984], for example, the first stage is a standard prisoner's dilemma and in the second, each of the two is free to punish the other, at some cost to herself, no matter what the other's first round behavior. He observes that under reasonable parameter restrictions, there is one subgame perfect equilibrium, in which neither cooperates or punishes, and nine Nash equilibria in pure strategies. More important, he also shows when each of the pure strategies is identified with a sub-population of players, and the model is enhanced to include a ninth sub-population, a set of best responders blessed with perfect recognition, there are two evolutionary stable states (ESS) and one neutrally stable state (NSS). The first ESS is monomorphic: "venegeful cooperators," those who cooperate and punish defection, comprise the entire population. In the second, polymorphic, ESS, "bullies," or defectors who punish other defectors, and best responders co-exist in fixed proportions. A continuum of population states, in which an indeterminate mixture of "passive defectors," those who defect and never punish, and best responders co-exist comprise the NSS. If the third is consistent with, and close(r) in spirit to, subgame perfection, the first is consistent with the existence of "strongly reciprocal norms," and simulation results based on the so-called replicator dynamic (Taylor and Jonker [1978]) suggest that both outcomes are locally stable, with non-negligible basins of attraction.

Sethi's [1996] model therefore provides a plausible rationale for strong reciprocity, but it remains to be seen whether *socially* reciprocal behavior can also be robust. The issue is more complicated than first seems, however, because the literature on ESS's in multi-player games that do not involve "playing the field" remains thin: in Broom, Cannings and Vickers' [1997] seminal paper, for example, even symmetric games exhibit "levels" of ESSs, and some of the canonical properties of pairwise models - the Bishop-Cannings Theorem, for example - do not generalize. More important, perhaps, even if contribution were evolutionarily stable in some weak(er) sense, there is reason to be concerned that, in the evocative language of Binmore and Samuelson [1999], such outcomes would comprise a "hanging valley" vulnerbale to the presence of "drift."

Following Binmore and Samuelson [1999], we shall instead consider the evolution of pure strategies under a perturbed selection mechanism - that is, a selection mechanism subject to drift - as the amount of drift tends to zero. Our own approach is unusual, however, inasmuch as we provide behavioral microfoundations for *both* the selection and drift functions. In particular, we suppose that there are five subpopulations - free riders, second order free riders, strong reciprocators, pure social reciprocators and social reciprocators - associated with each of the five pure strategies and that within each of these, there exist two sorts of "reinforcement-based learners," one more common *and* more sophisticated than the other. The more sophisticated are assumed to "sample and imitate" à la Nachbar [1990], in which case the unperturbed selection mechanism assumes the form of a scaled replicator dynamic, as confirmed below. The less sophisticated, on the other hand, are assumed to be aspiration-driven learners of the sort described in Carpenter and Matthews [2001], where the difference reflects how available information is, or is not, used, either in the lab or the outside world: imitation requires that some, if not all, of the available information be processed, while our version of the aspiration parable does not.

To be more precise, 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 Δ of the entire population re-evaluates its current performance.[12] A fraction $1-\theta$

[12] If instead we had assumed that the fraction was $k\Delta < 1$, the results would be identical, modulo a rescaled time variable.

of these are assumed to "sample" another member of the population - that is, observe or perhaps be told their behavior and outcome - and to switch, and therefore imitate, whenever (a) the sampled payoff is higher, and (b) the difference in payoffs exceeds some switching cost c, the value of which is a random variable with uniform distribution over $[0, \bar{c}]$. To ensure that the likelihood of a switch is always less than or equal to one, it is further assumed that $\bar{c} \geq 67.5$. The remaining θ percent do not sample but rather compare their current outcome to some *aspiration level* a, the value of which is also assumed to be a random variable, with uniform distribution over $[0, \bar{a}]$, where $\bar{a} \geq 87.5$. If it equals or exceeds this aspiration, the agent does not change her behavior, but if it falls short, she switches to another. In standard aspiration-based models (Binmore, Gale and Samuelson [1995], for example) the likelihoods that behaviors will be adopted are equal to their current population shares, but this assumes (a) the shares are observed and processed and, perhaps more important, (b) that the dissatisfied will sometimes switch back to their initial choices, which seems implausible to us. Modifying the "no switch back model" in Carpenter and Matthews [2001], we shall instead suppose that those who fall short of their aspirations will switch to one of the *other* strategies with equal likelihood.

Under these assumptions, the share p_i of the population committed to behavior i will evolve as follows:

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

The second term, for example, is the net increase in the share of i attributable to imitation. Of the $(1-\theta)\Delta p_i$ percent of the population committed to i who re-evaluate their performance each period, a fraction $p_j \max[0, \pi_j - \pi_i]$ will sample someone committed to $j \neq i$ who did better. Given the determination of switching costs, it follows, therefore, that a fraction $(1-\theta)\Delta p_i \bar{c}^{-1} p_j \max[0, \pi_j - \pi_i]$ of population will switch from i to $j \neq i$ as the result of sophisticated reinforcement, and that the total number of "defections" from i to all $j \neq i$ will be $(1-\theta)\Delta p_i \bar{c}^{-1} \sum_{j\neq i} p_j \max[0, \pi_j - \pi_i]$. In a similar vein, a fraction $(1-\theta)\Delta p_i \bar{c}^{-1} \sum_{j\neq i} p_j \max[0, \pi_i - \pi_j]$ of the population will switch from $j \neq i$ to i each period as a result of imitation.

The third term is the net increase in the share of sub-population i due to unsophisticated learning: the likelihood that someone who is committed to $j \neq i$ falls short of her/his aspiration level is $(\bar{a} - \pi_j)/\bar{a}$, so that a fraction $\theta\Delta\bar{a}^{-1}\sum_{j\neq i} p_j(\bar{a} - \pi_j)$ of the population will be dissatsified with $j \neq i$, a quarter (0.25) of whom will then switch to i, and so on.

Since the second term collapses to $\pi_i - \bar{\pi}$, where $\bar{\pi}$ is the expected population-wide payoff, (8.1) can be rewritten as:

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

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

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

and, in the special case of no drift or $\theta = 0$:

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

which is a scaled replicator dynamic.

Our principal concern here is the behavior of (8.3), but a simple simulation exercise based on (8.4) serves to underscore some basic themes. We first observe that, consistent with intuition, the expected payoff to each of the four classes of contributors will be a function of the proportion p_1 of (first order) free riders alone:

$$\begin{aligned} \pi_2 &= 75 - 37.5p_1 \\ \pi_3 &= 75 - 47.5p_1 \\ \pi_4 &= 75 - 57.5p_1 \\ \pi_5 &= 75 - 62.5p_1 \end{aligned} \tag{8.5}$$

Given the costs of of punishment, it comes as no suprise, for example, that for fixed p_1, second order free riders do better than strong reciprocators, who in turn do better than pure social reciprocators, or that social reciprocators will be the least successful of the four. What is important to understand, however, is that social reciprocators, pure or otherwise, need not vanish because first order free riders will sometimes do (much) worse since:

$$\pi_1 = 27.5 + 22.5p_1 + 60p_2 + 40p_3 + 20p_4 \tag{8.6}$$

after substitution for $\pi_5 = 1 - p_1 - p_2 - p_3 - p_4$.

Consider, for example, the case of an initial "balanced population," in which $p_i = 0.20$ for all i. The representative second order free rider will receive $75 - 37.5(0.20) = 67.5$ EMUs, the strong reciprocator, 65.5, the pure social reciprocator, 63.5 and the social reciprocator, 62.5, but the mean for first order free riders is (just) 56, with a population mean of 63. Given the rules of imitation, some first order free riders and social reciprocators will become second order free riders, a smaller number will become strong reciprocators, and a still smaller number will become pure social reciprocators. The first order free riders are more vulnerable, however - the likelihood that the difference in outcomes will exceed the costs of switching is greater - so that it is at least *possible* that the non-contributors will be driven to "extinction" before the social reciprocators, in which case the "fitness differential" $\pi_i - \bar{\pi}$ across contributors will vanish. Indeed, simulation of the RD from an initial balanced population reveals that (in rounded numbers) $p_1 \rightarrow 0$, $p_2 \rightarrow 0.34$, $p_3 \rightarrow 0.26$, $p_4 \rightarrow 0.22$ and $p_5 \rightarrow 0.18$.

This outcome is consistent with the behavior in at least some of our sessions, but it remains to be seen whether the prediction is robust with respect to drift. We are more interested, in other words, in the case where θ is small than when it is 0. The problem, of course, is that closed form solutions to (8.3), expressed as a function of θ, are difficult to obtain. Instead, we first computed (with Maple) the rest points of (8.3) for four different values of the drift parameter, $\theta = 0.10, 0.01, 0.001$ and 0.0001, and for $\bar{a} = \bar{c} = 100$. The results, together with the eigenvalues for the relevant Jacobians, are reported in Table 5, and derived in Appendix D.

[Insert Table 5 About Here]

Neglecting the case $\theta = 0.10$ for a moment, there are three such rest points, the characteristics of which are robust with respect to the amount of drift. In the first, there are almost no first order free riders, and the proportions of second order free riders and strong reciprocators are (to two decimal places) 32 and 26 percent, respectively, with smaller, and almost equal, numbers of the two sorts of social reciprocators. The second is consistent with standard game theoretic predictions: the proportion of first order free riders increases from 97.7 percent to 99.9 as θ falls from 0.01 to 0.0001, the proportion of second order free riders falls from 1.0 to 0.1 percent, and the proportions of all three sorts of reciprocators are smaller still. The third is similar to the first in the sense that there are few first order free riders, but differs inasmuch as more than half, as opposed to a third, of the population are first order free riders.

The case of $\theta = 0.10$ is different, however, because there is so much drift that no sub-population is ever able to dominate: in the second sort of equilibrium, for example, the proportion of first order free riders is less than 65 percent.

As the eigenvalues in Table 5 also indicate, however, while the first and second equilibria are locally stable, the third is not. Figures 6, 7, 8a and 8b illustrate some of the possible paths. Figure 6, for example, plots the evolution of shares from an initial "balanced population" - that is, $p_i = 0.20$ for all i - for the benchmark case $\theta = 0.01$, $\bar{a} = \bar{c} = 100$. Under these conditions, the shares rapidly converge to the first equilibrium, in which more than two thirds of the population will contribute and punish those who do not, and in which more than three fifths of this sub-population are social reciprocators of one kind or another. Furthemore, the proportions are close, if not equal, to those obtained in the first simulation exercise with the same initial condition but no drift.

[Insert Figures 6, 7, 8a and 8b About Here]

What ensures that the (almost) all contribution outcome will be stable? It is useful to decompose the selective pressures that exist at this point. In the case where $\theta = 0.01$ - so that $p_1 = 0.004632$, $p_2 = 0.318817$, $p_3 = 0.258326$, $p_4 = 0.217130$ and

$p_5 = 0.201095$ - the normalized fitness differentials $p_i(\pi_i - \bar{\pi})$ are:

$$\begin{aligned}
p_1(\pi_1 - \bar{\pi}) &= 0.004632(61.408880 - 74.708783) = -0.000616 \\
p_2(\pi_2 - \bar{\pi}) &= 0.318817(74.826300 - 74.708783) = +0.000375 \\
p_3(\pi_3 - \bar{\pi}) &= 0.258326(74.779980 - 74.708783) = +0.000184 \\
p_4(\pi_4 - \bar{\pi}) &= 0.217130(74.733660 - 74.708783) = +0.000054 \\
p_5(\pi_5 - \bar{\pi}) &= 0.201095(74.710500 - 74.708783) = +0.000003
\end{aligned}$$

In the absence of drift, 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 number of first order free riders to start with, their decrease is substantial. Furthermore, of these, a little more than 60% would become second order free riders and another 30 % would become strong reciprocators.

So what then prevents evolution toward these two behaviors, which would leave the population vulnerable to an influx of first order free riders? It is the behavior of aspiration-driven learners which provides the required "offset." To see this, observe that the drift terms are:

$$\begin{aligned}
0.25\sum_{j\neq 1} p_j(\bar{a} - \pi_j) - p_1(\bar{a} - \pi_1) &= 6.278116 - 0.178754 = +6.099361 \\
0.25\sum_{j\neq 2} p_j(\bar{a} - \pi_j) - p_2(\bar{a} - \pi_2) &= 4.316353 - 8.025803 = -3.709450 \\
0.25\sum_{j\neq 3} p_j(\bar{a} - \pi_j) - p_1(\bar{a} - \pi_3) &= 4.694057 - 6.514987 = -1.820929 \\
0.25\sum_{j\neq 4} p_j(\bar{a} - \pi_j) - p_1(\bar{a} - \pi_4) &= 4.951284 - 5.486080 = -0.534796 \\
0.25\sum_{j\neq 5} p_j(\bar{a} - \pi_j) - p_1(\bar{a} - \pi_5) &= 5.051406 - 7.760481 = -0.034186
\end{aligned}$$

The numbers confirm that first order free riders are both the one sub-population to lose from imitation *and* the one to benefit from dissatisfaction. Furthermore, of the four sorts of contributors, second order free riders lose the most "crude learners." To understand this, we observe that while the likelihood $(100 - 61.408880/100 \approx 0.386$ or 38.6 %) that the representative first order free rider will fall short of her aspiration level exceeds that of all four other sub-populations, there are so few first order free riders to start with that the absolute number of "defections" will be small. On the other hand, the likelihoods that crude contributors will become disenchanted are smaller and similar in size - from 25.2% for second order free riders to 25.3% for social reciprocators - but because all four, in particular second order free riders, are much more numerous, the number of defections is much higher. Furthermore, because one quarter of all the disenchanted contributors will "experiment" with non-contribution, first order free riders benefit most. Second order free riders, on the other hand, are hurt most because more of these switch from, and fewer switch to, this behavior. Because just one percent of the population are assumed to be crude learners, these forces cancel one another.

Viewed from another perspective, the assumed nature of drift in this model implies that at this (equilibrium) point, there is a constant source of new first order free riders. Because these first order free riders can expect to earn much less in a world where almost all others are contributors, however, there is also a constant, and equal, stream of defections.

Figure 7, on the other hand, plots the evolution of population shares from the unbalanced initial condition in which first order free riders comprise half the population ($p_1 = 0.50$), second order free riders another 20 percent ($p_2 = 0.20$), and strong, pure social and social reciprocators 10 percent each ($p_3 = p_4 = p_5 = 0.10$). In this case, there is rapid and almost monotone convergence to the no contribution equilibrium.

Figures 8a and 8b, 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, $p_1 = 0.02$, $p_2 = 0.54$, $p_3 = 0.21$, $p_4 = 0.13$ and $p_5 = 0.10$, and Figure 8a plots the evolution of population shares over the same time horizon as Figures 6 and 7, a horizon that was sufficient to ensure convergence in the first two cases. There is an almost imperceptible movement in the structure of the population, with a small increase in the share of second order free riders at the expense of first order free riders, from which one is tempted to infer the existence of a plateau of sorts. Figure 8b, which provides a (very) long run perspective on the same dynamics, reveals that such a conclusion would be premature: in short order, the proportion of first order free riders explodes, the proportion of second order free riders collapses, and the shares settle down, once and for all, at the second, no contribution, equilibrium. In this case, the model exhibits what is almost a régime shift, from a scenario in which almost all of the players contribute to the public good to one in which almost none of them do. While we observed a collapse of this sort in one or two experimental sessions, these were also followed by a "rebirth" of the contribution norm.

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. This question does not arise when $\theta = 0$ because in the absence of drift, the choice of $\bar{c}$ influences the "speed" of population shares on their respective time paths but not the properties of these paths. The introduction of behavioral drift, however, precludes the use of normalized dynamics. Tables 6 and 7 present comparative statics for the model's two stable equilibria for alternative values of $\bar{a}$ and $\bar{c}$ in the benchmark case $\theta = 0.01$.

[Insert Tables 6 and 7 About Here]

The results show that when there is not much drift, the equilibrium shares are not much affected, even when $\bar{a}$ and $\bar{c}$ increase from 100 to 200. Furthemore, the changes are consistent with intuition. An increase in the value of $\bar{c}$, for example, causes the expected costs of switching to rise or, in more evocative terms, increases the amount of "inertia": to induce "low performers" to switch, the difference in outcomes must be more substantial. This in turn reduces (increases) the selective pressure on less (more) successful strategies, which implies that their equilibrium shares should be lower (higher), and this is what happens. In the first, contribution-driven, equilibrium, the proportions of all four sorts of contributors are smaller - the differences, however, are from the third decimal place onward - while the proportion of first order free riders more than doubles, from 0.46 percent to 0.91. In the second, the share of first order free riders falls more than 2 percent, from 97.7 percent to 95.2, while the shares of the four sorts of contributors rise a little bit.

In a similar vein, an increase in $\bar{a}$ increases the likelihood that an individual will fall short of her/his aspiration - that is, become dissastified - no matter how successful (in relative terms, at least) their approach to MSR, so that here, too, one would expect the shares of more successful strategies to decrease, and vice versa, and this is indeed the case.

To be consistent with our experimental data, however, contributors must survive for more than some small and perhaps contrived set of initial conditions. That is, the first equilibrium should have a substantial basin of attraction. Given the dimension of (8.3), however, the two basins are difficult to characterize in graphical terms, and some sort of "compression" is needed. To this end, Figure 9 plots the proportions of first and second order free riders for the initial conditions $p_1 = 0.1, 0.2, \ldots, 1$ and $p_i = (1 - p_i)/4$ for all $i \neq 1$ - that is, when the shares of the four sorts of contributors are equal. When the share of first order free riders is about a quarter or smaller of the population, almost all of these will be driven to abandon such behavior, but when the share exceeds this, there is convergence to the no contribution equilibrium. In some cases, however, convergence is slow, and exhibits the same sort of sudden shifts illustrated in Figure 8b. On the path marked AA, for example, which tends, in the long run, to the no contribution equilibrium, the number of first order free riders is still *falling* at $t = t_1$, when the other paths are close to their eventual rest points. Furthemore, when it happens, the "turnaround" on AA is rapid, as the movement between t_1 and t_2 illustrates.

Figure 10, on the other hand, plots the same proportions for the initial conditions $p_2 = 0.1, 0.2 \ldots, 1$ and $p_i = (1 - p_i)/4$ for all $i \neq 2$, so that the initial proportions of first order free riders and contributors who punish are equal. In this case, when the initial share of second order free riders is a third or less, first order free riders (almost) vanish, but when it is exceeds this threshold, there is (sometimes slow and roundabout) convergence to the no contribution equilibrium. What both Figures 9 and 10 reveal, however, is that the survival of reciprocal behavior, both strong and social, is not limited to a small neighborhood of initial conditions.

9. Conclusion

This paper provides an integrated - survey, experiment and model - perspective on "social reciprocity," which we define as the willingness to enforce norms regardless of 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" but also, on the basis of the second stable equilibrium, the observation that even villages can sometimes fall short.

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 - that is, cases in which reciprocators punish free riders rather than reward other contributors - 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 either sophisticated or crude, but even the sophisticated learners' rule is a simple one, and it remains to show whether our results can be extended to models with still more sophisticated cognition. It is possible, for example, that another set of rules would be consistent with both the sudden collapse of contribution norms, as ours are, but also with their sudden rebirth, as evidenced in some sessions of our experiment.

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.

References

Acheson, J. (1993). Capturing the commons: Legal and illegal strategies. *The political economy of customs and culture: Informal solutions to the commons problem.* T. Anderson and R. Simmons Eds. Lanham, Rowman and 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.

Banarjee, A. and J. Weibull. (1994). Evolutionary selection and rational behavior. *Rationality and Learning in Economics.* A. Kirman and M. Salmon Eds. Oxford, Basil Blackwell.

Binmore, K. G. and L. Samuelson. (1994). An economist's perspective on the evolution of norms. *Journal of Institutional and Theoretical Economics* 150: 45-63.

Binmore, K. G. and L. Samuelson. (1999). Evolutionary drift and equilibrium selection. *Review of Economic Studies* 66: 363-393.

Binmore, K. G., J. Gale and L. Samuelson. (1995). Learning to be imperfect: the ultimatum game. *Games and Economic Behavior* 8: 56-90.

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: The effects of residual claimancy and reciprocity. mimeo.

Bowles, S. and H. Gintis (1999). The evolution of strong reciprocity. *Santa Fe Institute Working Paper.*

Broom, M., C. Cannings and G. T. Vickers (1997). Multi-player matrix games. *Bulletin of Mathematical Biology* 59: 931-952.

Carpenter, J. (2001). Punishing free-riders: How group size affects mutual monitoring and collective action. *mimeo.*

Carpenter, J. and P. H. Matthews. (2001). No switchbacks: Rethinking aspiration-based dynamics in the miniature ultimatum game. *mimeo.*

Carpenter, J., P. H. Matthews and O. Ong'ong'a. (2002). Why punish? Social reciprocity and the enforcement of prosocial norms. *mimeo.*

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.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives* 3(4): 99-117.

Fehr, E. and S. Gaechter (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90(4): 980-994.

Ghemawat, P. (1995). Competitive advantage and internal organization: Nucor revisited. *Journal of Economics and Management Strategy* 3(4): 685-717.

Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206: 169-179.

Güth, W. and Kliemt, H. 1993. Competition or cooperation: on the evolutionary economics of trust, exploitation and moral attitudes, *Metroeconomica*, 45: 155-187.

Harrald, P. and Morrison, W. G. 2001. Sleepers, thresholds, and gateways: drift and stability in dynamic evolutionary games, *Wilfred Laurier University Working Paper*.

Isaac, R. M., J. Walker and S. Thomas (1984). Divergent evidence on free-riding: An experimental examination of possible explanations. *Public Choice* 43(1): 113-149.

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. *The handbook of experimental economics*. J. Kagel and A. Roth Eds. Princeton, Princeton University Press: 111-194.

Moir, R. (1998). Spies and swords: Costly monitoring and sanctioning in a common-pool resource environment. *mimeo*.

Mudd, S. (1968). Groups sanction severity as a function of degree of behavior deviation and relevance of norm. *Journal of Personality and Social Psychology* 8(3): 258-260.

Mudd, S. (1972). Group sanction severity as a function of degree of behavior deviation and relevance of norm: Replication and revision of model. *The Journal of Psychology* 80: 57-61.

Nachbar, I. (1990). Evolutionary selection in dynamic games. *International Journal of Game Theory* 19: 59-90.

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: 404-417.

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(August 15): 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: 113-140.

Sethi, R. and E. Somanathan (2001). Understanding reciprocity, *mimeo*, forthcoming, *Journal of Economic Behavior and Organization*.

Shotland, L. and M. Straw (1976). Bystander response to an assault: When a man attacks a woman. *Journal of Personality and Social Psychology* 34(5): 990-999.

Taylor, P. and Jonker, L. (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40: 145-156.

Appendix A: Survey Vignettes

STRONG RECIPROCITY: You and a number of other newly hired people are employed by an auto manufacturer and assigned to work in teams of four. Everyone on the team is paid equally and the pay level is determined entirely by how many cars your work team produces. On the first day of work, you and the other three members of your team divide up the production tasks equally. Over the course of the next month, you and two other members of your group work regularly and hard. However, the fourth member of the team often hides in a storage room and reads a book instead of working on cars. This means the other three of you must work harder to make the same number of cars as the other four-person teams. At the end of the month, you and everyone else in your group earn the same amount of money.

SOCIAL RECIPROCITY (LOW COST): You and a number of other newly hired people are employed by an auto manufacturer and assigned to work in teams of four. Everyone on the team is paid equally and the pay level is determined entirely by how many cars your work team produces. On the first day of work, you and the other three members of your team divide up the production tasks equally. Each of you works equally hard making cars. However, you notice that a member of the group that occupies the workspace next to yours often hides in a storage room and reads a book instead of working on cars. While your earnings are unaffected by what this member of the other team is doing, the members of his team must work harder and share their income with this person.

SOCIAL RECIPROCITY (HIGH COST): You and a number of other newly hired people are employed by an auto manufacturer and assigned to work in teams of four. Everyone on the team is paid equally and the pay level is determined entirely by how many cars your work team produces. On the first day of work, you and the other three members of your team divide up the production tasks equally. Each of you works equally hard making cars. However, you notice that a member of the group that occupies the workspace next to yours often hides in a storage room and reads a book instead of working on cars. This person sees that you have noticed that he is not working. He approaches you and says, "If you tell anyone about this, I will do something bad to you." While your earnings are unaffected by what this member of the other team is doing, the members of his team must work harder and share their income with this person.

Appendix B: 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 $5.00 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:

$$1 \times (\text{EMUs you keep}) + 0.5 \times (\text{Total EMUs contributed by your group})$$

Lets 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 $4 \times 15 = 60$ EMUs. Each group member earns $0.5 \times 60 = 30$ 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 woul

Example 2: Now say everyone in the group contributes 5 EMUs. Here the group total contribution will be 20 and each member will earn $0.5 \times 20 = 10$ 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 $3 \times 25 = 75$ EMUs. Each group member earns $0.5 \times 75 = 37.5$ 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 persons earnings into the input-text box that appears below that participants 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 dont 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 persons earnings will be reduced by 6 EMUs. If the other participants spend 4 EMUs in total, the persons 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.

Nobodys 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:

$$\begin{aligned}&\text{(Gross Earnings from Stage One)} - (2 \times \text{the number of EMU spent}\\ &\qquad\qquad\qquad\qquad\qquad\quad \text{on reductions directed towards you)}\\ &\qquad\qquad\qquad\qquad\qquad - \text{(your expenditure on reductions directed}\\ &\qquad\qquad\qquad\qquad\qquad\quad \text{at other participants)}\end{aligned}$$

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

Appendix C. 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 $\sigma^i = (0, p_2, p_3, p_4, p_5)$ for all $i = 1, \ldots, 4$. 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 $\sigma^j = (p_1^j, p_2^j, p_3^j, p_4^j, p_5^j)$ where $p_1^j > 0$, with payoff $\pi^j(\sigma^j, \sigma^i, \sigma^i, \sigma^i)$. It follows that $\pi^j = p_1^j \pi_1^j + (1 - p_1^j)75 = 75 + p_1^j(\pi_1^j - 75)$, where π_1^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, $\pi_1^j < 75$. 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 $p_2(p_2+p_3)^2 + p_4(p_2+p_3)^2 = (p_2+p_4)(p_2+p_3)^2$, where the first term is the product of the likelihood p_2 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 $2p_2(p_2+p_3)(p_4+p_5)+p_3(p_2+p_3)^2+2p_4(p_2+p_3)(p_4+p_5)+p_5(p_2+p_3)^2$, 47.5 with likelihood $p_2(p_4+p_5)^2+2p_3(p_2+p_3)(p_4+p_5)+p_4(p_4+p_5)^2+2p_5(p_2+p_3)(p_4+p_5)$, and 27.5 with likelihood $p_3(p_4+p_5)^2 + p_5(p_4+p_5)^2$. Gathering terms, we have:

$$\begin{aligned}
\pi_1^j = {} & 87.5p_2(p_2+p_3)^2 + 87.5p_4(p_2+p_3)^2 + 135p_2(p_2+p_3)(p_4+p_5) \\
& + 67.5p_3(p_2+p_3)^2 + 135p_4(p_2+p_3)(p_4+p_5) + 67.5p_5(p_2+p_3)^2 \\
& + 47.5p_2(p_4+p_5)^2 + 95p_3(p_2+p_3)(p_4+p_5) + 47.5p_4(p_4+p_5)^2 \\
& + 95p_5(p_2+p_3)(p_4+p_5) + +27.5p_3(p_4+p_5)^2 + 27.5p_5(p_4+p_5)^2
\end{aligned}$$

or, after factoring:

$$\begin{aligned}
\pi_1^j &= (p_2+p_4)[87.5(p_2+p_3)^2 + 135(p_2+p_3)(p_4+p_5) + 47.5(p_4+p_5)^2] \\
&\quad (p_3+p_5)[67.5(p_2+p_3)^2 + 95(p_2+p_3)(p_4+p_5) + 27.5(p_4+p_5)^2] \\
&= (p_2+p_4)[87.5(p_2+p_3) + 47.5(p_4+p_5)][p_2+p_3+p_4+p_5] \\
&\quad (p_3+p_5)[[67.5(p_2+p_3) + 27.5(p_4+p_5)][p_2+p_3+p_4+p_5]
\end{aligned}$$

Since $p_2 + p_3 + p_4 + p_5 = 1$, this can be rewritten:

$$\begin{aligned}
\pi_1^j &= (p_2+p_3)(87.5(p_2+p_3) + 67.5(p_3+p_5)) \\
&\quad + (p_4+p_5)(47.5(p_2+p_4) + 27.5(p_3+p_5)) \\
&= 87.5(p_2+p_4) + 67.5(p_3+p_5) - 40(p_4+p_5) \\
&= 87.5p_2 + 67.5p_3 + 47.5p_4 + 27.5p_5
\end{aligned}$$

It follows, therefore, that $\pi_1^j < 75$ if and only if:

$$87.5p_2 + 67.5p_3 + 47.5p_4 + 27.5p_5 < 75$$

or, since $p_2 = 1 - p_3 - p_4 - p_5$ in this case:

$$20p_3 + 40p_4 + 60p_5 > 12.5$$

or:

$$p_3 + 2p_4 + 3p_5 > 0.625$$

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 $\sigma^i = (p_1, 1 - p_1, 0, 0, 0)$: 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.

Appendix D. The Stability Properties of MSR's Rest Points

Since the domain of (8.3), the four dimensional simplex, is invariant, we shall limit attention to the first first four equations and substitute for $p_5 = 1 - p_1 - p_2 - p_3 - p_4 - p_5$. Rewriting (8.3) as $\dot{p} = (1-\theta)f(\cdot) + \theta g(\cdot)$, the $(i,j)^{th}$ element of the Jacobian F of the selection function $f(\cdot)$ is:

$$\begin{aligned} f_{i,i} &= \bar{c}^{-1}\left((\pi_i - \bar{\pi}) + p_i \frac{\partial(\pi_i - \bar{\pi})}{\partial p_i}\right) \\ f_{i,j} &= \bar{c}^{-1} p_i \frac{\partial(\pi_i - \bar{\pi})}{\partial p_j} \qquad\qquad j \neq i \end{aligned}$$

The fact that $\partial\pi_i/\partial p_j = 0$ for i and $j \neq 1$ simplifies the calculation of F, so that:

$$F = \bar{c}^{-1}\begin{pmatrix} \alpha_1 & p_1(60 - 85p_1) & p_1(40 - 55p_1) & p_1(20 - 25p_1) \\ p_2(\alpha_2 - 85p_1) & p_1(\alpha_2 - 85p_2) & -55p_1p_2 & -25p_1p_2 \\ p_3(\alpha_2 - 85p_1 - 10) & -85p_1p_3 & p_1(\alpha_2 - 10 - 55p_3) & -25p_1p_3 \\ p_4(\alpha_2 - 85p_1 - 20) & -85p_1p_4 & -55p_1p_4 & p_1(\alpha_2 - 20 - 25p_4) \end{pmatrix}$$

where

$$\begin{aligned} \alpha_1 &= (-47.5 + 275p_1 + 60p_2 + 40p_3 + 20p_4) \\ &\quad - p_1(255p_1 + 170p_2 + 110p_3 + 50p_4) \\ \alpha_2 &= 72.5 - 85p_1 - 85p_2 - 55p_3 - 25p_4 \end{aligned}$$

The $(i,j)^{th}$ element of the Jacobian G of the drift function $g(\cdot)$ is likewise:

$$\begin{aligned} g_{i,i} &= 0.25\bar{a}^{-1}[(\pi_5 - \bar{a}) - \sum_{k\neq i} p_k \frac{\partial \pi_k}{\partial p_i)} - (1 - \sum_k p_k)\frac{\partial \pi_5}{\partial p_i}] - 1 + \bar{a}^{-1}[p_i \frac{\partial \pi_i}{\partial p_i} + \pi_i] \\ g_{i,j} &= 0.25\bar{a}^{-1}[(\pi_5 - \pi_i) - \sum_{k\neq i} p_k \frac{\partial \pi_k}{\partial p_j} - (1 - \sum_k)\frac{\partial \pi_5}{\partial p_j} + \bar{a}^{-1} p_i \frac{\partial \pi_i}{\partial p_j} \end{aligned}$$

which leads, after some simplification, to:

$$g_{i,j} = \bar{a}^{-1}\begin{pmatrix} \alpha_3 - 1.25\bar{a} & 53.75p_1 & 36.25p_1 & 18.75p_1 \\ \alpha_4 - 46.875p_2 & \alpha_5 - 1.25\bar{a} & -13.75p_1 & -6.25p_1 \\ \alpha_4 - 47.5p_3 & -21.25p_1 & \alpha_5 - 5p_1 - 1.25\bar{a} & -6.25p_1 \\ \alpha_4 - 71.875p_4 & -21.25p_1 & -13.75p_1 & \alpha_5 - 10p_1 - 1.25\bar{a} \end{pmatrix}$$

where

$$\begin{aligned} \alpha_3 &= 61.875 + 13.75p_1 + 53.75p_2 + 36.25p_3 + 18.75p_4 \\ \alpha_4 &= 27.5 - 42.5p_1 - 21.25p_2 - 13.75p_3 - 6.25p_4 \\ \alpha_5 &= 93.75 - 68.125p_1 \end{aligned}$$

The $(i,j)^{th}$ element of the Jacobian J of the perturbed selection mechanism is therefore $(1-\theta)f_{i,j} + \theta g_{i,j}$.

For example, in the case where $\theta = 0$ - that is, there is no drift - the Jacobian evaluated at $p_1 = 0, p_2 = p_3 = p_4 = 0$ is:

$$J = \begin{pmatrix} -0.375 & -0.250 & -0.150 & -0.050 \\ 0 & -0.125 & 0 & 0 \\ 0 & 0 & -0.225 & 0 \\ 0 & 0 & 0 & -0.325 \end{pmatrix}$$

J is triangular, so the diagonal entries, $-0.375, -0.125, -0.225$ and-0.325, are the eigenvalues. Since all of these are negative and real, the no contribution equilibrium is stable in the absence of noise, consistent with intuition. If $\theta = 0$ but $p_1 = 0$, on the other hand, the Jacobian becomes:

$$J = \begin{pmatrix} -0.475 + 0.600p_2 + 0.400p_3 + 0.200p_4 & 0 & 0 & 0 \\ p_2(0.725 - 0.850p_2 - 0.550p_3 - 0.250p_4) & 0 & 0 & 0 \\ p_3(0.625 - 0.850p_2 - 0.550p_3 - 0.250p_4) & 0 & 0 & 0 \\ p_4(0.525 - 0.850p_2 - 0.550p_3 - 0.250p_4) & 0 & 0 & 0 \end{pmatrix}$$

with eigenvalues $-0.475 + 0.600p_2 + 0.400p_3 + 0.200p_4$ and 0, repeated three times. When $p_3 + 2p_4 + 3p_5 > 0.625$, the abovementioned condition for a SNE, is satisfied, the first of these is negative, but the repeated zero eigenvalues preclude a definitive characterization.

With the introduction of aspiration-driven learners or "noise," however, the properties of the three rest points are not difficult to characterize. In the case $\theta = 0.01$, for example, the Jacobian J associated with the first equilibrium, $p_1 = 0.004632$, $p_2 = 0.318817$, $p_3 = 0.258326$ and $p_4 = 0.217130$, is equal to:

$$J = \begin{pmatrix} -0.131021 & 0.002758 & 0.001839 & 0.000921 \\ 0.078902 & -0.003236 & -0.000810 & -0.003684 \\ 0.038637 & -0.001017 & -0.003106 & -0.000299 \\ 0.010700 & -0.000856 & -0.000554 & -0.003164 \end{pmatrix}$$

for which the eigenvalues, -0.133335, -0.001695, -0.003000, -0.002487, are all real and negative. Likewise, for the second equilibrium, $p_1 = 0.976659$, 0.010281, 0.005700, 0.003943, the Jacobian is:

$$J = \begin{pmatrix} -0.336621 & -0.217290 & -0.129080 & -0.040871 \\ -0.011132 & -0.132340 & -0.006810 & -0.003096 \\ -0.007375 & -0.006760 & -0.224100 & -0.001988 \\ -0.005943 & -0.005316 & -0.003440 & -0.319199 \end{pmatrix}$$

and this, too, has real and negative eigenvalues, -0.121120, -0.361670, -0.312956 and -0.216488. For the third equilibrium, $p_1 = 0.019156$, 0.537293, 0.207271, 0.128402, the Jacobian is:

$$J = \begin{pmatrix} -0.195352 & 0.011173 & 0.007455 & 0.003738 \\ 0.046322 & -0.009908 & -0.005631 & -0.002559 \\ -0.001950 & -0.003382 & -0.005315 & -0.000995 \\ -0.013791 & -0.002111 & -0.001366 & -0.005777 \end{pmatrix}$$

and while three of the eigenvalues are once more real and negative, -0.037672, -0.003307 and -0.005400, the fourth is real and positive, 0.005954.

	Strong Reciprocity	Social Reciprocity (Low Cost)	Social Reciprocity (High Cost)
The shirker should be punished.	97	96	96
I would confront the shirker.	90	43	51

Notes: The sample size *n* was 79, and responses were given on a seven point Likert scale.

Table 1. Why Punish? Material Harm or Norm Violation
(Percent of Participants Responding Affirmatively)

Dependent Variable = $Punishment_{ij}$
(All regressions include random effects.)

	(1)	(2)	(3)	(4)
	All Data	All Data	Ingroup	Outgroup
$Constant$	0.08	0.36***	-0.08	0.06
	(0.13)	(0.12)	(0.22)	(0.04)
$FreeRide$	1.74***	–	1.93**	0.20**
	(0.19)		(0.80)	(0.09)
$FRDummy$	–	1.29***	0.82	0.31***
		(0.16)	(0.71)	(0.08)
$Outgroup$	-0.38***	-0.38***	–	–
	(0.13)	(0.13)		
$Strong$	0.25	0.23	–	–
	(0.22)	(0.22)	-	-
Adjusted R^2	0.02	0.02	0.02	0.04
Wald χ^2	104.08	80.98	29.84	77.12

Notes: Standard errors in parantheses. *** denotes significance at the 0.01 level, ** at the 0.05 level, and * at the 0.10 level.

Table 2. Why Punish? Material Harm or Norm Violation

Average Payoff Reduction Due To Punishment For Free Riders

	Period	1	2	3	4	5	6	7	8	9	10
Social Reciprocity	Free Rider < Group Average	6.22	5.72	8.00	7.90	12.78	5.21	8.62	5.34	13.34	26.66
	Free Rider < 1/3 Endowment	7.70	11.60	12.00	14.26	21.26	6.00	8.34	5.60	21.10	46.00
Strong Reciprocity	Free Rider < Group Average	6.40	9.60	12.22	11.26	7.34	3.76	6.28	10.80	7.14	22.00
	Free Rider < 1/3 Endowment	8.86	7.60	10.34	6.76	6.76	3.76	9.66	10.80	6.28	15.60

Wilcoxon Test	$z = 1.98, p = 0.05$
Kolmogorov-Smirnov Test	$KS = 0.18, p = 0.03$

Table 3. How Severely Are Free Riders Punished?

Dependent Variable = $PublicContribution_t$
(All regressions are tobit and include random effects.)

	(1)	(2)	(3)
$Constant$	8.35***	6.97***	8.35***
	(2.01)	(1.77)	(1.85)
$Public_{t-1}$	0.74***	0.70***	0.70***
	(0.09)	(0.09)	(0.09)
$Punishment_{t-1}$	0.45***	0.43***	-0.24
	(0.15)	(0.15)	(0.29)
$Social$	–	2.99***	-0.91-
		(1.47)	(1.81)
$Social\times$ $Punishment_{t-1}$	–	–	0.87***
			(0.31)
Wald χ^2	63.91	81.26	86.74

Notes: Standard errors in parantheses. *** denotes significance at the 0.01 level, ** at the 0.05 level, and * at the 0.10 level.

Table 4. Do Free Riders Respond To Punishment?

	$\theta = 0.10$	$\theta = 0.01$	$\theta = 0.001$	$\theta = 0.0001$
First Order Free Riders	0.044554	0.004632	0.000464	0.000046
Second Order Free Riders	0.295840	0.318817	0.321176	0.321411
Strong Reciprocators	0.247643	0.258326	0.259423	0.259532
Pure Social Reciprocators	0.212951	0.217130	0.217587	0.217633
Social Reciprocators	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
First Order Free Riders	0.649904	0.976659	0.999772	0.999773
Second Order Free Riders	0.158916	0.010281	0.001003	0.001000
Strong Reciprocators	0.084318	0.005700	0.000557	0.000056
Pure Social Reciprocators	0.057382	0.003943	0.000386	0.000038
Social Reciprocators	0.0494788	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
First Order Free Riders	0.308332	0.019156	0.001857	0.000185
Second Order Free Riders	0.337351	0.537293	0.551069	0.552410
Strong Reciprocators	0.159996	0.207271	0.209367	0.209565
Pure Social Reciprocators	0.104866	0.128402	0.129233	0.129311
Social Reciprocators	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

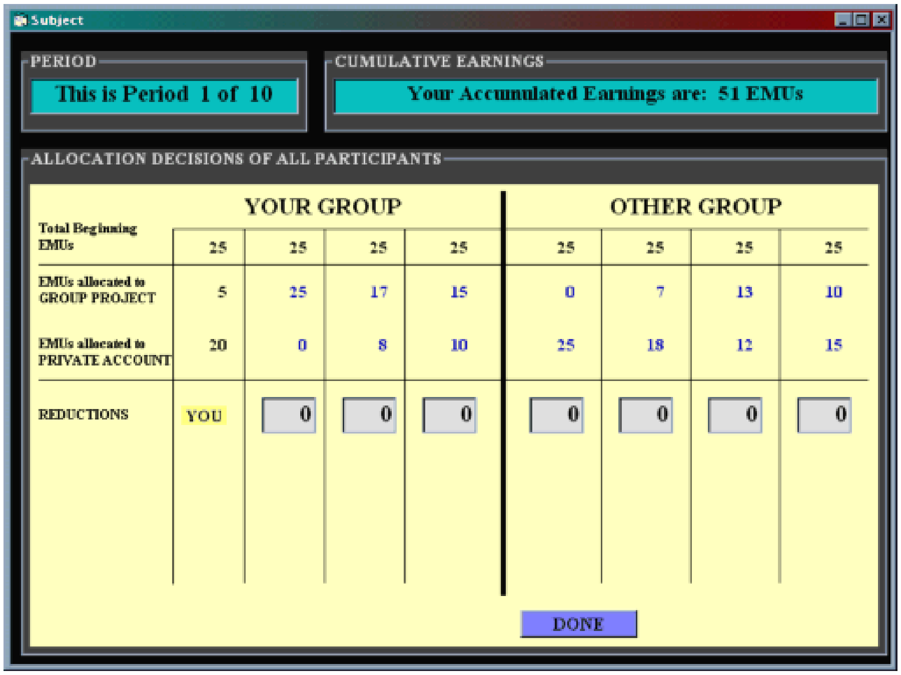
Table 5. Rest Points and Eigenvalues for MSR

	$\bar{c} = 100$	$\bar{c} = 150$	$\bar{c} = 200$
First Order Free Riders	0.004632	0.006909	0.009159
Second Order Free Riders	0.318817	0.317525	0.316247
Strong Reciprocators	0.258326	0.257726	0.257133
Pure Social Reciprocators	0.217130	0.216882	0.216638
Social Reciprocators	0.201095	0.200958	0.200824
First Order Free Riders	0.976659	0.964687	0.952496
Second Order Free Riders	0.010281	0.015567	0.020958
Strong Reciprocators	0.005700	0.008621	0.011593
Pure Social Reciprocators	0.003943	0.005961	0.008013
Social Reciprocators	0.003416	0.005164	0.006941

Table 6. The Comparative Statics of Switching Costs

	$\bar{a} = 100$	$\bar{a} = 150$	$\bar{a} = 200$
First Order Free Riders	0.004632	0.009172	0.011406
Second Order Free Riders	0.318817	0.316239	0.314967
Strong Reciprocators	0.258326	0.257129	0.256538
Pure Social Reciprocators	0.217130	0.216636	0.216395
Social Reciprocators	0.201095	0.200824	0.200693
First Order Free Riders	0.976659	0.968470	0.964291
Second Order Free Riders	0.010281	0.013895	0.015741
Strong Reciprocators	0.005700	0.007698	0.008717
Pure Social Reciprocators	0.003943	0.005323	0.006028
Social Reciprocators	0.003416	0.004612	0.005222

Table 6. Comparative Statics of Dissatisfaction



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

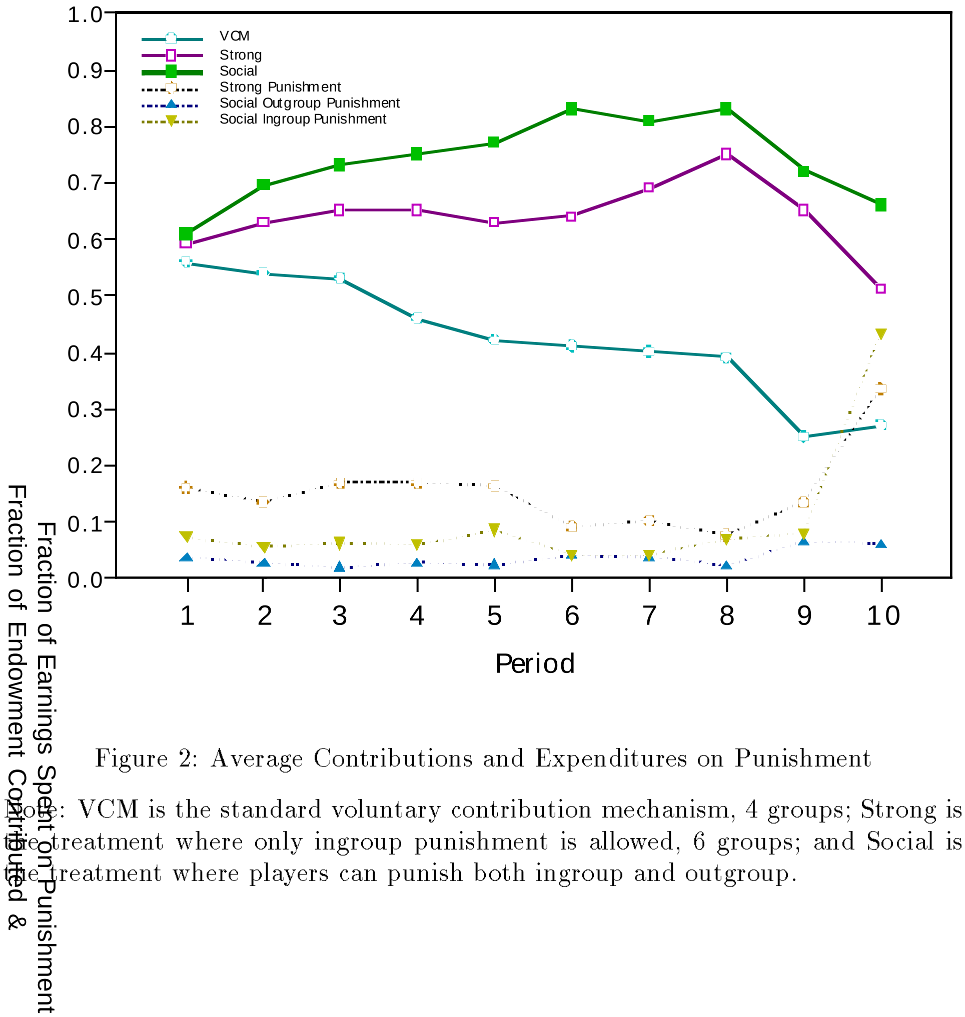


Figure 2: Average Contributions and Expenditures on Punishment

Note: VCM is the standard voluntary contribution mechanism, 4 groups; Strong is the treatment where only ingroup punishment is allowed, 6 groups; and Social is the treatment where players can punish both ingroup and outgroup.

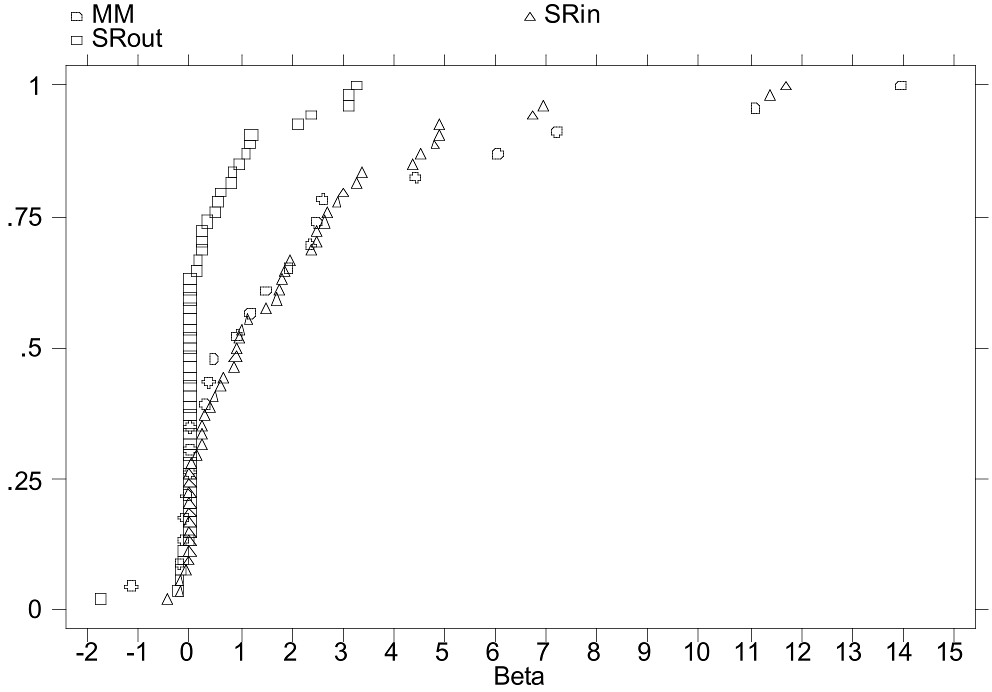


Figure 3: Cumulative Distribution of Punishment Coefficients - Propensities to Punish

Note: These are cumulative distributions of beta coefficients on the regression of punishment assigned on degree ot target's free riding. MM indicates the Strong Treatment, SRin is ingroup punishment in the Social treatment, and SRout is outgroup punishment in the Social treatment.

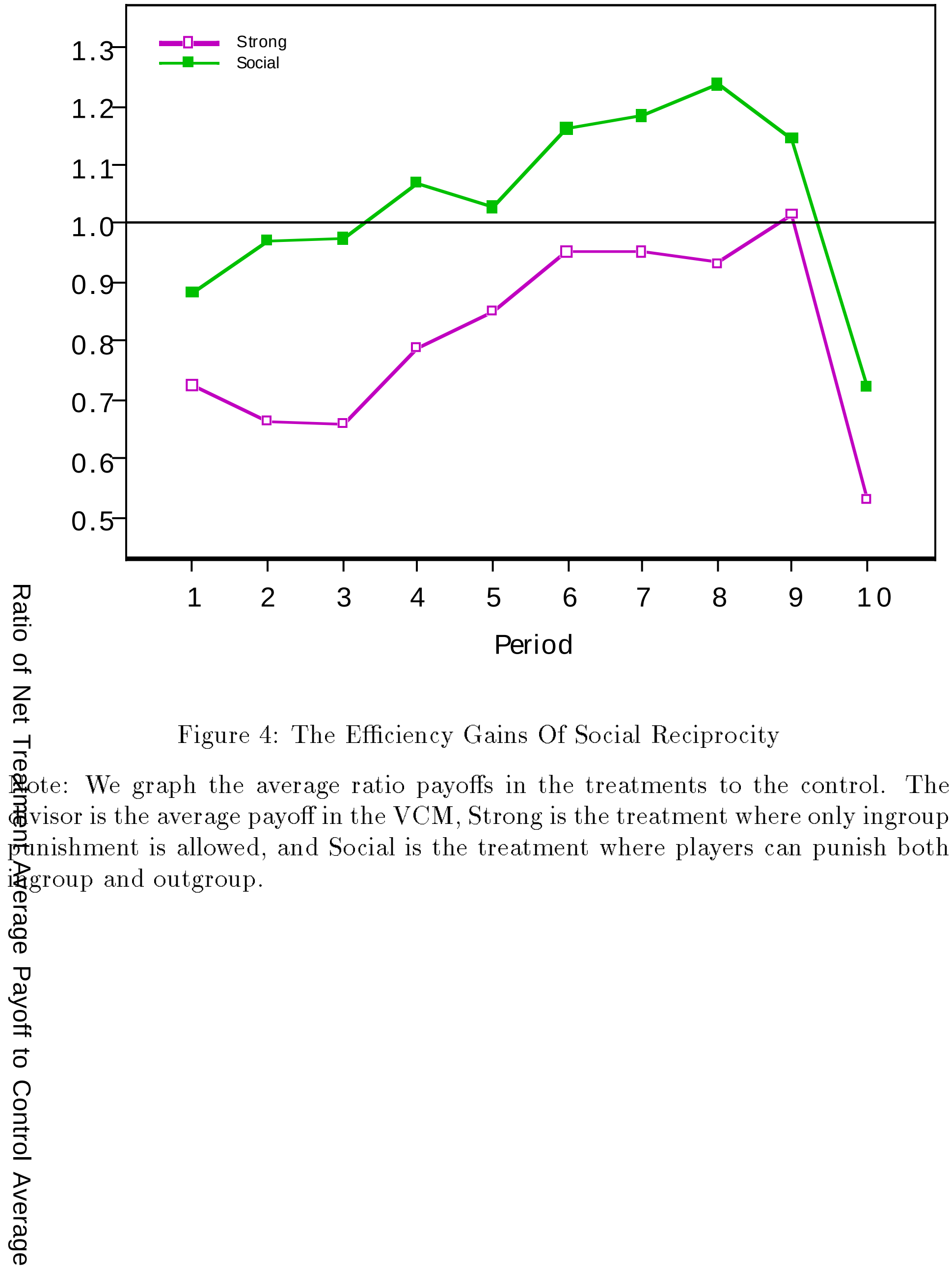


Figure 4: The Efficiency Gains Of Social Reciprocity

Note: We graph the average ratio payoffs in the treatments to the control. The divisor is the average payoff in the VCM, Strong is the treatment where only ingroup punishment is allowed, and Social is the treatment where players can punish both ingroup and outgroup.

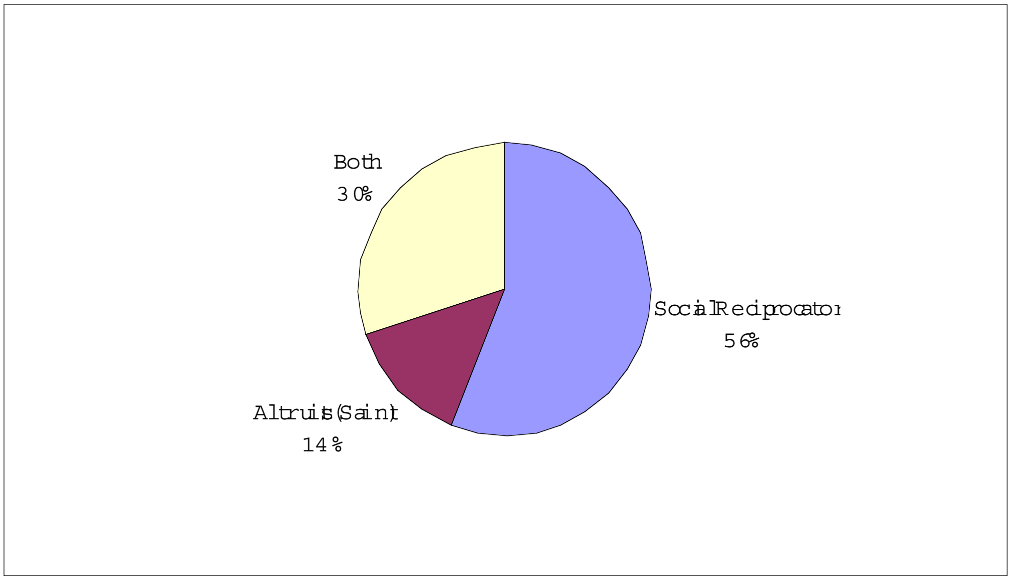


Figure 5: States Reasons For Outgroup Punishment

Note: 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 questions.

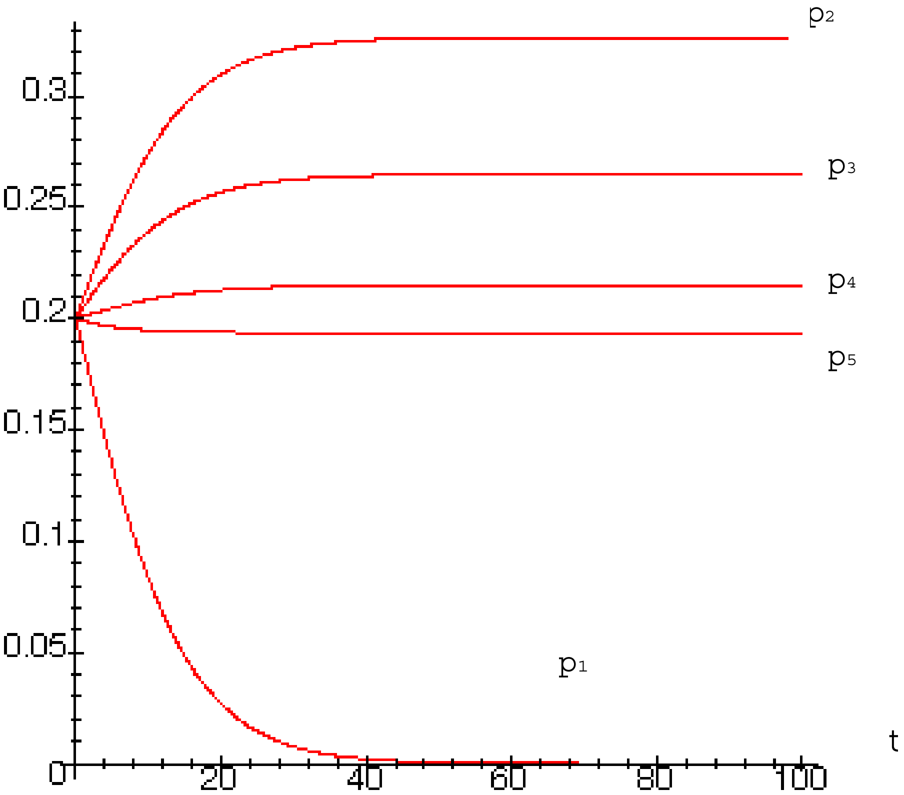


Figure 6: Evolution from an Initial Balanced Population

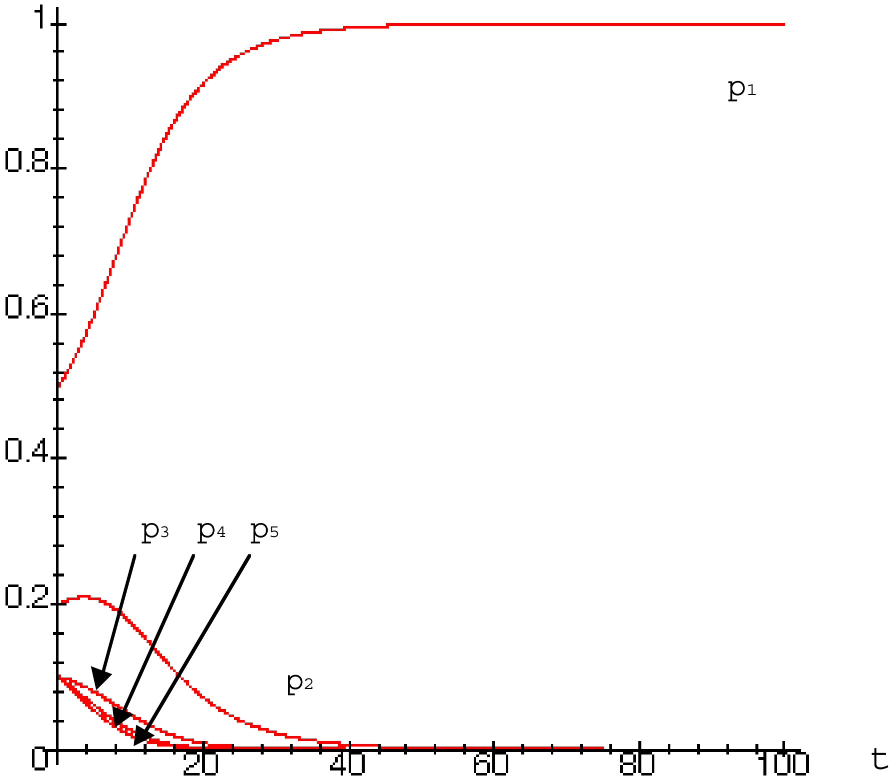


Figure 7: Almost Monotone Evolution To The No Contribution Equilibrium

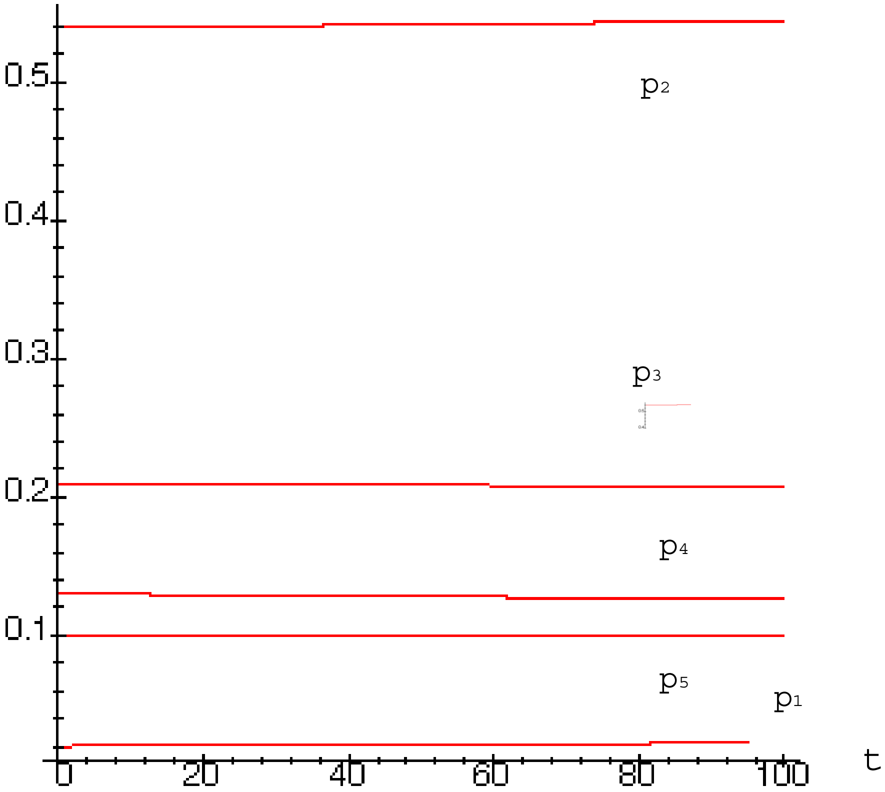


Figure 8a: A Plateau Near The Unstable Equilibrium

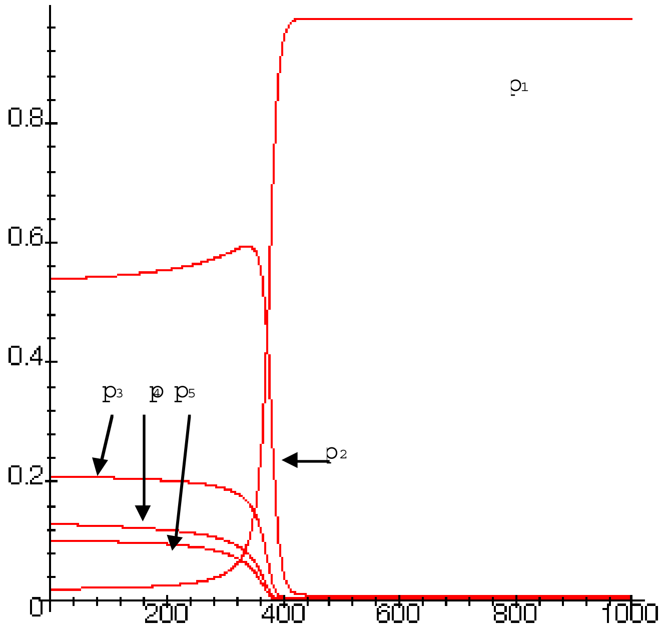


Figure 8b: "Falling Off The Plateau": The Long Run Instability of The Third Rest Point

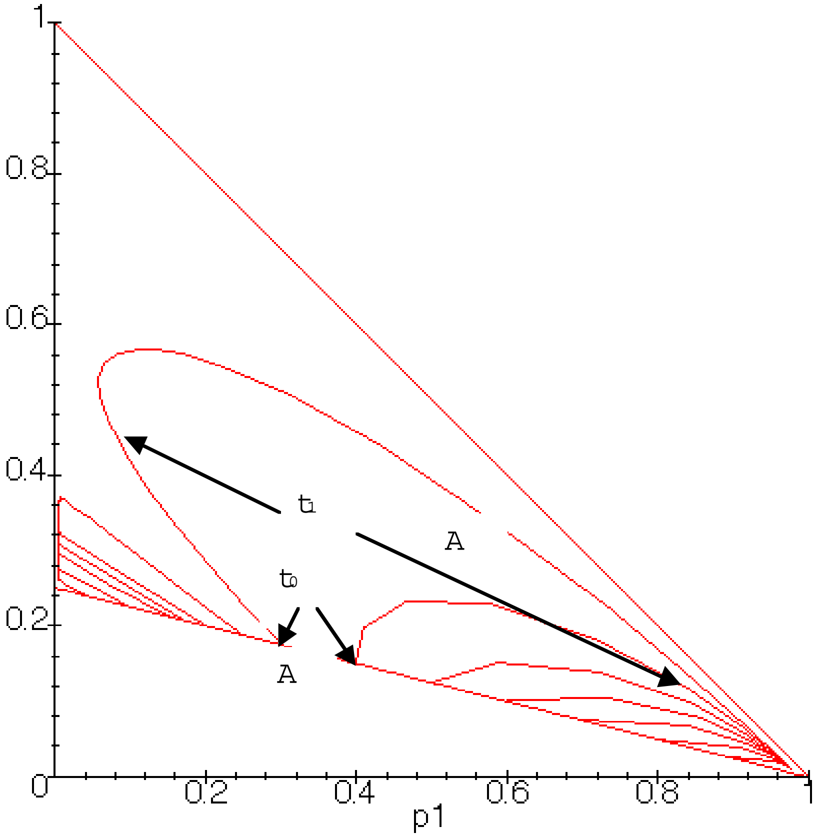


Figure 9: Basins of Attractions - First View

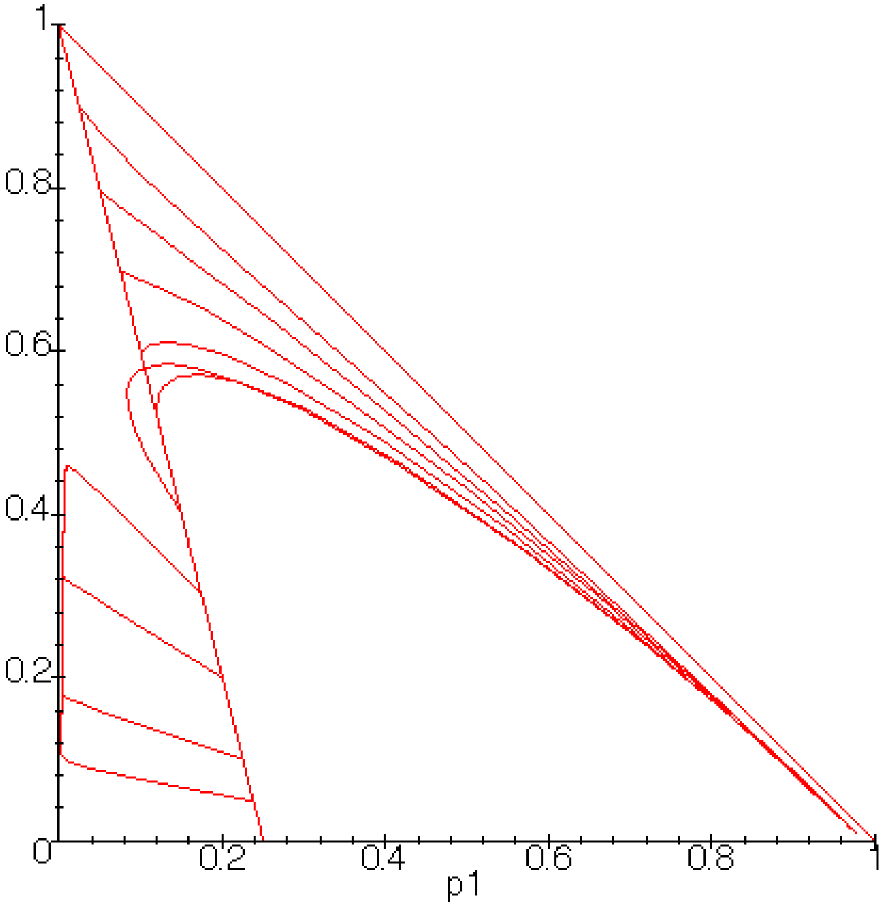


Figure 10: Basins of Attraction - Another View