Norm Enforcement: The Role of Third Parties

by

JEFFREY P. CARPENTER AND PETER HANS MATTHEWS*

> "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 [1751/1983, p. 93]

To be effective, norm enforcement often requires the participation of unaffected third parties. The logic of third-party intervention has, however, proven elusive because the costs always seem to outweigh the benefits. Using an evolutionary game theoretic approach, we posit that the intervention of unaffected bystanders is a triggered normative response and show that generalized punishment norms survive in one of the two stable equilibria subject to selection drift. (JEL: C 73, D 03, D 64, H 41)

1 Introduction

In 1964 the murder of Kitty Genovese in the courtyard of a Queens, New York housing complex shocked everyone including a number of social scientists. Aside from the brutal nature of the crime, one of the most shocking details was that nobody intervened to help. Casual observation suggests that it seems natural to many people to help others even when the costs are substantial and there are no obvious benefits to tip the scales. Indeed, the subsequent research of psychologists and sociologists on "bystander intervention," much of it motivated by the Genovese case, provides some support for this natural inclination. BOROFSKY, STOLLAK, AND MESSE [1971] and SHOTLAND AND STRAW [1976], for example, demonstrated that a significant number of people will intervene in a seemingly severe altercation between two people even though the one to intervene is not being harmed, nor is there any reason to expect that the one to intervene will receive any payoff from doing so. In the former, 29% intervened in situations in which two confederates of the experimenter staged an altercation that escalated into a physical fight. The latter found that a much higher proportion, 65%, intervened when a male confederate pretended to assault

* Middlebury College, VT, and Institute for the Study of Labor (IZA) (corresponding author) and Middlebury College, VT.

Journal of Institutional and Theoretical Economics
JITE 166 (2010), 239–258 © 2010 Mohr Siebeck – ISSN 0932-4569

a female confederate, but also that this number dwindled to 19% when the two confederates seemed to be married and the cost of intervening seemed higher.[1]

Intervention and punishment have also proven to be important in social dilemmas, for example, the provision of local public goods, common pool resource extraction or team production. Despite strong incentives to free-ride on the efforts of others, the members of groups who confront such dilemmas are sometimes adept at attenuating incentive problems without external intervention. Communities often develop rules that make contributing and free-riding transparent (OSTROM [1992]), but perhaps more importantly community members are also often willing to incur costs to monitor and punish behavior that benefits the individual but harms the group (e.g., ACHESON [1988]). Acts of this kind tend to maintain or increase the efficiency of social interactions so one might posit that monitoring is in the interests of group members. Indeed, it may be the case that if free-riders respond by contributing more in the future, the benefits that accrue to monitoring and punishing may exceed the individual costs, measured perhaps in terms of possible retaliation. However, even in this case we know that punishers can do better by free-riding on the punishment meted out by other monitors. In fact, by the same logic that not contributing dominates contributing to a public good in one-shot interactions (OLSON [1965]), there is no logic by which narrowly self-interested agents monitor and punish.

In this paper we are interested in understanding the origins, limits, and social implications of individuals who incur costs to express their disapproval of antisocial behavior. Our focus is on norm-driven reciprocity and, in particular, on the willingness of individuals to punish behaviors *both* when the punisher him/herself has been harmed and when neither the punisher, nor his/her group, has been harmed. Little or no attention has been paid to the latter, a form of "third-party punishment," in the literature, with the notable exceptions of BENDOR AND MOOKHERJEE [1990], BENDOR AND SWISTAK [2001], and FEHR AND FISCHBACHER [2004]. Bendor and Swistak, whose analysis is the closest to our own, provide the logic for how third-party punishment might evolve within groups in which repeated interactions are the norm. We consider our analysis to be an important extension of their pathbreaking work. We relax the assumption of repeated interactions by allowing agents to be randomly rematched between game cycles, we explicitly model the dynamics of evolution to reveal the possibility of states that hover nearby steady states in the medium run but quickly diverge after some time, we allow various punishing types to compete "head-to-head," and we show how punishment networks need not overlap with payoff networks. Third-party punishers in our model intervene when someone breaks the contribution norm in a completely separate group. In fact, one interpretation of our results suggests that as long as punishment networks are more comprehensive than payoff networks, norms will be maintained more effectively.

---

[1] As suggested by our referee, another important experiment in this line of research was conducted by DARLEY AND BATSON [1973] who showed that not being in a hurry was a better predictor of stopping to help a person in need than having just been asked to give a short talk on the parable of the Good Samaritan.

Fehr and Fischbacher find strong evidence of third-party punishment in their three-person dictator experiment, but suggest that our public-goods oriented design "allows for reciprocity and strategic interactions among third parties ... [so that we] cannot rule out third party punishment for reasons of self-interest" (FEHR AND FISCHBACHER [2004, p. 66]). Under our assumptions, however, agents never know who had, or had not, punished whom, so we are confident that self-interest is *not* an explanation.

To make our analysis more comprehensive, we distinguish between two types of norm-driven behavior based on group boundaries. *Strong Reciprocators* (CARPENTER et al. [2009], BOWLES AND GINTIS [2004], GINTIS [2000], SETHI [1996]) punish those members of their ingroup that free-ride, where an ingroup is loosely defined as the subset of individuals who benefit from a specific public good that they can all contribute to. *Social Reciprocators*, on the other hand, punish free-riders even in groups to which they can neither contribute to nor benefit directly from. In other words, they are unaffected third parties. Social reciprocity differs from strong reciprocity because social reciprocators punish all norm violators, regardless of group affiliation and with little regard to the social distance between punisher and norm violator, as long as there exists some "punishment network" that connects them. Further, while the trigger for punishment by strong reciprocators is the cost implicitly imposed by a free-rider on the group, we hypothesize that the trigger for social reciprocity is simpler. Social reciprocators just punish anyone who violates a contribution norm, and need not be harmed directly by the free-rider.[2]

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

The psychological experiments on bystander intervention mentioned above offer two more examples of the behavioral type that, for the sake of exposition we will continue to call social reciprocity, and a third can be found in LATANE AND DARLEY's [1970] work. In their experiment, subjects are asked to wait in a room to be interviewed. A confederate, also in the room, steals what remains of the show-up fee fund when the experimenter leaves. Their dependent variable is the probability that subjects report the theft when the experimenter returns. Because all subjects have been paid their show-up fee and therefore suffer no loss when the theft occurs, strong reciprocity is not an issue. Furthermore, since there is no expectation of a reward, there can be no instrumental reason for intervening, not least because the costs of turning in the confederate could be high. Despite this,

[2] See CARPENTER AND MATTHEWS [2010] for experimental evidence of this behavior.

in 50% of the cases in which the subjects reported noticing the theft, they turned in the confederate.

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

2 *Towards A Model of Third-Party Norm Enforcement*

Consider a "miniature social reciprocity game" (hereafter, MSR) consistent, in broad terms, with the incentives of many social dilemmas. In this case the underlying game is constructed to have the same incentives as the experimental participants who played a similar game in CARPENTER AND MATTHEWS [2010]. Suppose that, at each moment in discrete time, "nature" chooses a "punishment network" of four individuals at random from a large (technically, infinite) population and then divides each foursome into pairs. MSR is then played in two stages. In the first, each of the two pairs plays its own public-goods or voluntary contribution game, in which individuals must decide whether to contribute all or none of their endowment of 50 "utils" to a common fund with a marginal per capita return of fifty percent (i.e., contributions are summed, the total is multiplied by 1.5 and shared equally). The normal form of this version of the prisoner's dilemma played by each pair in the first stage is therefore:

Table 1
The Miniature Social Reciprocity Game

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

In the second stage, the choices of *all four* are then revealed to *all four*, after which *contributors* must decide (a) whether or not to enforce a "contribution norm" and punish free-riders and, if so, (b) which free-riders – ingroup, outgroup, or both – to punish. We suppose, for purposes of simplification, that those who punish outsiders, the social reciprocators, cannot "pick and choose." A contributor, for example, who is also committed to "norm enforcement" both within and across pairs and who

is matched with three – one in and two out – free-riders must sanction all three. Each punishment act is assumed to cost a contributor 10 EMUs, and to reduce a free-rider's payoff by 20 EMUs.

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

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

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

Punishment *is* observed, however, and it cannot be rationalized as either conditional cooperation or "strong reciprocity" in the sense of BOWLES AND GINTIS [2004]. On the one hand, because the foursomes are dissolved at the end of each period, no individual is ever matched, absent a measure zero coincidence, with someone from a previous foursome of his or hers. Such punishment cannot be understood, therefore, in terms of the Folk Theorem or the so-called "trigger strategies" that support conditional cooperation in some environments. On the other hand, the fact that at least some of this punishment is inflicted on outsiders implies that it cannot all be attributed to strong reciprocity, as CARPENTER, MATTHEWS, AND ONG'ONG'A [2004] have underscored.

Within the framework of the model, then, the question of whether some, or even all, of the continuum of all-contribute SNEs could meet some other, perhaps less restrictive, requirements for equilibrium becomes critical. In particular, we are interested in whether what we have called social reciprocity is, in a well-defined sense, *evolutionarily stable*. We should therefore first note that, as we have formalized it, MSR is an extension of what AXELROD [1984] first called the "Norms Game," a framework since featured in the research of GÜTH AND KLIEMT [1993], BINMORE AND SAMUELSON [1994], and SETHI [1996]. In SETHI's [1996] reformalization, two players confront the usual prisoner's dilemma, after which each is free to punish the other, at some cost to him/herself, no matter what the other's first-stage behavior. He then demonstrates that when each of the (eight) pure strategies available is identified with a sub-population of possible players, and a ninth sub-population, a set of best responders blessed with perfect recognition, is added, there will be two evolutionarily stable states (ESS) and one neutrally stable state (NSS).

It is the first, monomorphic, ESS, in which "vengeful cooperators" comprise the entire population, that is most relevant here, not least because simulation exercises, based on the replicator dynamic (TAYLOR AND JONKER [1978]), indicate that this outcome will be locally stable and that its basin of attraction could be substantial.

If SETHI's [1996] model provides a plausible explanation of the evolution of *strong* reciprocity in some environments, it remains to be seen whether *social* reciprocity can also sometimes survive selection pressures. Our approach here is not based on the ESS criterion (like BENDOR AND SWISTAK [2001]) – indeed, it is not clear how ESS should be defined in this context (BROOM, CANNINGS, AND VICKERS [1997]) – but rather the distinct notion of *drift compatible* population states (BINMORE AND SAMUELSON [1999]). Our implementation is also unique, however, because we provide microfoundations for *both* the selection mechanism and drift function in terms of "learning" or "cultural transmission."

To this end, suppose that there are now five subpopulations associated with each of the five pure strategies in MSR and, in a convenient abuse of notation, denote their respective shares p_F, p_C, p_I, p_O, and p_B. To further streamline the exposition, we shall refer to their respective members as free-riders, second-order free-riders, strong reciprocators, pure social reciprocators, and social reciprocators. The evolution of population shares over time is then assumed to reflect two sorts of *reinforcement-based learning*, one more sophisticated and more common than the other. We suppose that sophisticated learners "sample and imitate" in the sense of NACHBAR [1990], in which case the selection mechanism assumes the form of a scaled replicator dynamic, as confirmed below. The less sophisticated, on the other hand, are aspiration-driven, as described in CARPENTER AND MATTHEWS [2005], where the difference reflects how available information is, or is not used.

To be more precise, we suppose for the moment that time is marked in discrete intervals of length Δ and that at the end of each of these periods, a fraction $k\Delta$ of the entire population re-evaluates their present situations. A proportion $1 - \theta$ of these, where θ is small, will sample another member of the population – that is, observe or somehow learn their behavior and outcome – and to switch or imitate (a là SCHLAG [1998]) whenever (a) the sampled payoff is higher and (b) the difference exceeds some switch cost c, the value of which is a random variable with uniform distribution over $[0, \bar{c}]$. To ensure the likelihood of a switch is always less than or equal to one, it is further assumed that $\bar{c} \geq 67.5$. A proportion θ, on the other hand, compare their current situation to some *aspiration level* a, the value of which is also a random variable, with uniform distribution over $[0, \bar{a}]$, where $\bar{a} \geq 87.5$. If one's payoff equals or exceeds this aspiration, the individual does not switch, but if it falls short, he or she "experiments" with another behavior. In the standard aspiration model (BINMORE, GALE, AND SAMUELSON [1995], for example), the probabilities that behaviors are adopted are assumed equal to their current population shares, but this implies that (a) these shares are observed and this information is processed and, more important, (b) the dissatisfied will sometimes "switch back" to the behavior that produced the dissatisfaction, neither of which seems desirable to us. Instead, we shall use a modified "no switch back dynamic" (CARPENTER AND MATTHEWS

[2005]), where individuals who have fallen short of their aspirations are assumed to switch to *another* pure strategy at random. It is this behavior that produces "drift" or "mutation" in our model.

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

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

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

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

Since the bracketed expression in the second term collapses to the measure of "differential fitness" $\pi_i - \bar{\pi}$, where $\bar{\pi}$ is the average payoff for the population as a whole, (1) can be rewritten as

$$
\begin{aligned}
\frac{p_i(t+\Delta) - p_i(t)}{\Delta} &= (1-\theta)k\bar{c}^{-1}p_i(\pi_i - \bar{\pi}) \\
&+ \theta k\bar{a}^{-1}\left[0.25\sum_{j\neq i} p_j(\bar{a}-\pi_i) - p_i(\bar{a}-\pi_i)\right].
\end{aligned} \tag{2}
$$

As $\Delta \to 0$, we have the continuous time version of (2):

$$
\dot{p}_i = (1-\theta)\bar{c}^{-1}p_i(\pi_i - \bar{\pi}) + \theta\bar{a}^{-1}\left[0.25\sum_{j\neq i} p_j(\bar{a}-\pi_i) - p_i(\bar{a}-\pi_i)\right] \tag{3}
$$

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

In the special case where there is no drift ($\theta = 0$) or aspiration-driven "mutation," (3) is the standard replicator dynamic:

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

While our principal concern here is the behavior of (3), a brief discussion of the evolution of shares in the absence of drift provides some important intuition. We first note that the expected payoffs for the four subpopulations of contributors are a function of p_F, the proportion of first-order free-riders, alone:

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

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

$$\pi_F = 27.5 + 22.5 p_F + 60 p_C + 40 p_I + 20 p_O$$

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

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

third of the population will contribute but not punish, but 40 (= 22 + 18) percent will be social reciprocators of one kind or another.

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

There is reason to be concerned, however, that the all-contribute equilibria of (3) are vulnerable to random drift. It should be noted, however, that while it is not difficult to posit some "mutation" – the *massive* and *simultaneous* transformation of all kinds of contributors into first-order free-riders, for example – that would undo such equilibria, shocks of this sort are implausible. Rather, the issue here is whether or not the existence of small but persistent "noise" will push the population far from this component and toward the all free-ride equilibrium. We are especially interested, for example, in whether outcomes in which all four contribute constitute a "hanging valley" (BINMORE AND SAMUELSON [1999]) that is consistent with medium-run equilibrium, a topic we return to when discussing an unstable interior equilibrium of the dynamics. In mechanical terms, our focus is on the behavior of (3) as θ tends to zero.

Closed form solutions to (3), expressed as a function of the drift paramater θ, are difficult (if not impossible) to obtain, however, so we report computed (with Maple) solutions for three values of θ, 0.01, 0.001, and 0.0001, with the relevant eigenvalues, in Table 2, for the case in which $\bar{a} = \bar{c} = 100$.

Table 2 reveals that under (3), MSR has *three* rest points, the properties of which seem robust with respect to the amount of drift. We are confident, in other words, that the compositions of the population in the limit, as $\theta \to 0$, are close to these. In the first, there are almost no free-riders – in rounded numbers, the proportion is 0.4% when $\theta = 0.01$, and falls to 0.004% when $\theta = 0.0001$ – and the share of second-order free-riders, those who contribute but do not enforce norms, is about 32% in all three cases. Most important from the perspective of both our experimental results and model, however, almost 42% of the population are social reciprocators of one kind or another, and are therefore prepared to punish outsiders who do not contribute. This is, therefore, our "reciprocal equilibrium."

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

The third is similar to the first in the sense that there are almost no first-order free-riders, but there are also fewer social reciprocators – in each case, a little less than 24% – and more second-order free-riders. As Table 2 also reveals, however,

Table 2
Rest Points and Eigenvalues for MSR

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

this equilibrium is not stable: three of the four eigenvalues are negative, but the fourth is positive. The fact that it is also small, however, has important implications, as seen below.

Figures 1 through 4 illustrate some possible solution paths. Figure 1, for example, plots the evolution of shares from a position of initial balance – that is, $p_i(0) = 0.20$ for all i – for the benchmark case $\theta = 0.01$, $\bar{a} = \bar{c} = 100$. As in the case of no drift, the population converges, rapidly, to the all-contribute and the limit values are not far apart.

What forces ensure that this outcome is stable, despite the continuous re-introduction of first-order free-riders to the population? It is useful to decompose the

selective pressures that exist in this case. In the benchmark case, the normalized fitness differentials are

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

In the absence of mutation, then, the representative first-order free-rider does much worse than all four sorts of contributors, each of whom receives more than the population mean, so much so that despite the small size of their subpopulation, the decrease in their numbers is a substantial one. On the other hand, more than 60% of the free-riders who switch as a result of imitation will become contributors who (also) do not punish and another 30% will become contributors who do not punish outsiders.

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

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

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

Figure 1
Evolution from an Initially Balanced Population

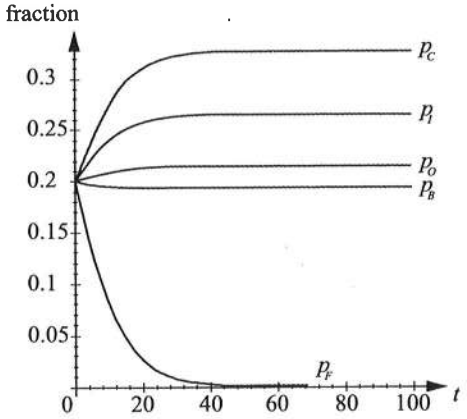
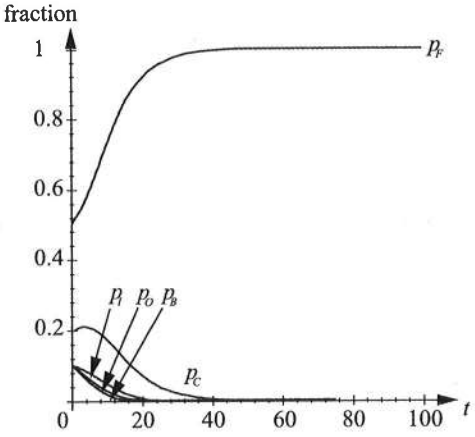


Figure 2
Almost Monotone Evolution to the No-Contribution Equilibrium



because more switch from, and few switch to, this behavior. Because the proportion of aspiration-based learners is just one percent, these cancel one another out.

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

Figure 2 depicts the evolution of population shares from the unbalanced initial condition in which first-order free-riders comprise half the population ($p_F = 0.50$),

Figure 3
A Plateau Near the Unstable Equilibrium

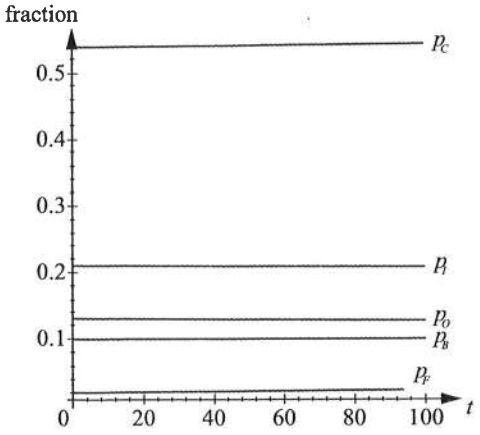
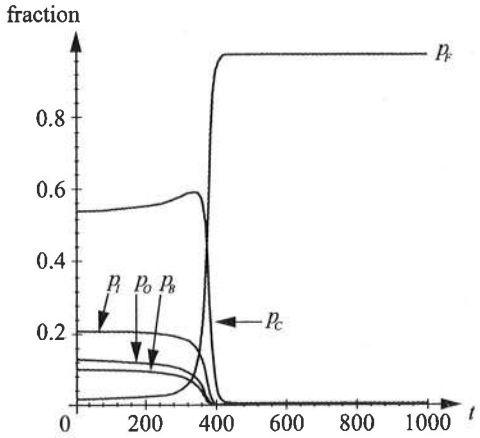


Figure 4
Falling off a Plateau: The Long-Run Instability of the Third Equilibrium



second-order free-riders another 20% ($p_C = 0.20$), and strong, pure social and social reciprocators 10% each ($p_I = p_O = p_B = 0.10$). In this case, there is rapid and almost monotone convergence to the no-contribution equilibrium.

Figures 3 and 4, on the other hand, illustrate one of the more "exotic" possibilities that follow from the introduction of drift. The initial point is chosen close to the third, unstable, equilibrium, $p_F = 0.02$, $p_C = 0.54$, $p_I = 0.21$, $p_O = 0.13$, and $p_B = 0.10$, and Figure 3 plots the evolution of population shares *over the same time horizon* as Figures 1 and 2, a period of time more than sufficient to "settle down" in those cases. It seems that there is an almost imperceptible drift in the population, *from* first-order

free-riders *toward* second-order free-riders, and perhaps a plateau of sorts. Figure 4, which provides a much longer-run perspective on the same dynamics, demonstrates that this conclusion would be premature. Indeed, Although the fraction of second-order free-riders appears to have stabilized in the time frame allowed in Figure 3, Figure 4 indicates that drift has pushed the population over the edge of a hanging valley indicated by the simultaneous collapse of all the cooperative types and the explosion of first-order free-riders. In the end, a stable no-contribution equilibrium is established. In this case, the model exhibits what is in effect a régime shift, from a situation in which almost all contribute to one in which almost no one does.

Given a fixed value of θ, each of the stable rest points is hyperbolic, so that small changes in the values of either $\bar{a}$ or $\bar{c}$ will have small changes on equilibrium shares, but it is important to ask what would happen if, for example, one of the parameters doubled in size. The issue is moot, of course, in the absence of drift, since aspiration levels are (in this case, at least) irrelevant and the switch cost affects the speed of evolution but not its path. To this end, Tables 3 and 4 present some comparative statics for the model's two stable equilibria.

The results show that when there is not much drift, the equilibrium shares are not much affected, even when the sizes of $\bar{a}$ and $\bar{c}$ double, from 100 to 200. Furthermore, the effects on the equilibrium shares are consistent with intuition. An increase in the value of $\bar{c}$, for example, increases the amount of "inertia": to induce the less successful to switch, the difference in outcomes must be more substantial. This in turn reduces the selective pressure on less successful behaviors, which implies that their equilibrium shares will decrease, and this is indeed what happens. In the reciprocity equilibrium, the proportions of all four sorts of contributors become smaller – the differences, however, are from the third decimal place onward – while the proportion of first-order free-riders increases, from 0.46% to 0.91%. For the

Table 3
The Comparative Statics of Switching Costs

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

Table 4
The Comparative Statics of Dissatisfaction

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

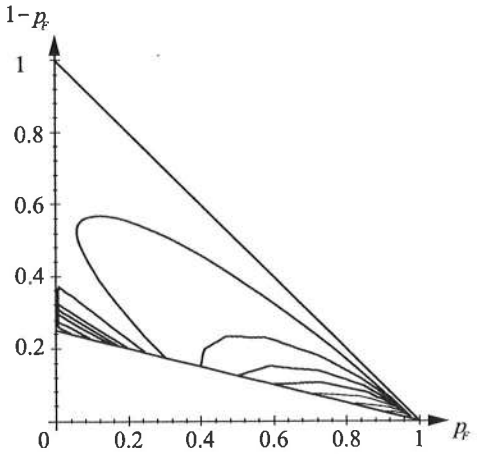
same reason, the share of first-order free-riders in the no-contribution equilibrium falls 2.5%, to 95.2%, while the shares of all four sorts of contributors increase a little bit.

In a similar vein, an increase in $\bar{a}$ increases the likelihood that an individual will fall short of his or her aspiration no matter how successful (in relative terms, at least) their MSR outcomes, so that here, too, one would expect the shares of "favored subpopulations" to decrease, and vice versa, and the results in Table 4 confirm this.

To be consistent with experimental data (e.g., CARPENTER AND MATTHEWS [2010]), however, it must also be the case that contributors survive under more than some small and perhaps contrived set of initial conditions. That is, the first equilibrium should be stable and have a substantial basin of attraction. Given the dimension of (3), a pictorial characterization is difficult, and some sort of compression is needed. To this end, Figure 5 plots the evolution of first- and second-order free-riders for the set of initial conditions $p_F(0) = 0.1, 0.2, \ldots, 1$ and $p_i(0) = (1 - p_F(0))/4$ for all $i \neq F$ – that is, for the case where the initial shares of the four sorts of contributors are equal. It shows that when the initial share of first-order free-riders is less than 25% or so, reinforcement-based "evolution" will drive the population to the all-contribute equilibrium, but when the initial share exceeds this, the population instead moves toward the no-contribution equilibrium. In some cases, the process is slow and exhibits the same sudden shifts illustrated in Figure 4: on the most dramatic of the inverted u-shaped paths, for example, there is a sudden turnaround in the fortunes of first-order free-riders at time t_1.

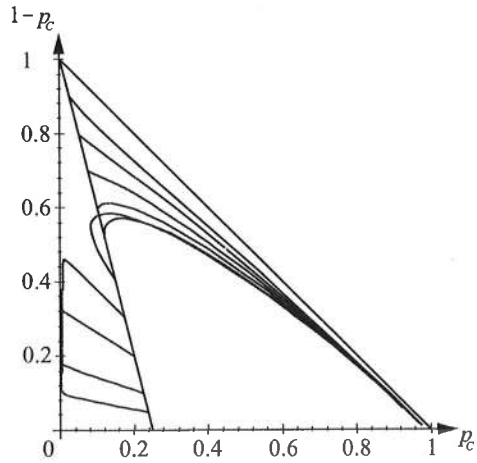
Figure 6 illustrates the evolution of the same two population shares under a different set of initial conditions, $p_C(0) = 0.1, 0.2, \ldots, 1$ and $p_i(0) = (1 - p_C(0))/4$ for all $i \neq C$, a condition that equalizes the numbers of free-riders and each of the three

Figure 5
A View of the Basins of Attraction



Note: p_F is the fraction of free-riders.

Figure 6
Another View of the Basins of Attraction



Note: p_C is the fraction of unconditional cooperators.

sorts of contributors who punish. In this case, when the initial share of second-order free-riders is a third or less, first-order free-riders almost vanish, consistent with the intuition that for non-contributors to flourish, the combined shares of those prepared to punish such behavior must be smaller than some threshold value. Otherwise, there is sometimes slow and roundabout evolution toward the no-contribution equilibrium.

The crucial common feature of Figures 5 and 6 is that the survival of reciprocity, both strong and social, is not unusual or limited to a small neighborhood of the all-contribute equilibrium.

3 Conclusion

This paper provides a theoretical perspective on the ability of third-party punishment to evolve and help maintain cooperative norms. 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 will sometimes fall short of the mark.

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

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

Appendix: Symmetric Nash Equilibria (SNE)

We shall first show that the two common profiles identified in the text are indeed SNEs, and then show that no others are possible. The argument that the first profile – that is, the case in which all four choose to free-ride – satisfies this criterion is trivial, so we shall focus on the second, in which all four randomize over the four pure contribution strategies. Consider the common mixture $\sigma^i = (0, p_C, p_I, p_O p_B)$ for all $i = 1, \ldots, 4$. 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 $\sigma^j = \left(p_F^j, p_C^j, p_I^j, p_O^j, p_B^j\right)$ where $p_F^j > 0$, with payoff $\pi^j(\sigma^j, \sigma^i, \sigma^i, \sigma^i)$. It follows that $\pi^j = p_F^j \pi_F^j + \left(1 - p_F^j\right)75 = 75 + p_F^j\left(\pi_F^j - 75\right)$, where π_F^j is what j can expect to earn as a unilateral free-rider, and therefore

that there will be no incentive to deviate from σ^i if $\pi^j < 75$ or, substituting in the previous expression, $\pi_F^j < 75$. Under what circumstances will this condition be met? That is, under what conditions can the unilateral free-rider expect to receive less than 75? We first observe that she will earn 87.5 with likelihood $p_C(p_C + p_I)^2 + p_O(p_C + p_I)^2 = (p_C + p_O)(p_C + p_I)^2$, where the first term is the product of the likelihood p_C that her partner will choose to contribute but not punish and the likelihood that both members of the outgroup will either contribute but not punish or contribute and punish insiders. Following similar logic, she will receive 67.5 with likelihood $2p_C(p_C + p_I)(p_O + p_B) + p_I(p_C + p_I)^2 + 2p_O(p_C + p_I)(p_O + p_B) + p_B(p_C + p_I)^2$, 47.5 with likelihood $p_C(p_O + p_B)^2 + 2p_I(p_C + p_I)(p_O + p_B) + p_O(p_O + p_B)^2 + 2p_B(p_C + p_I)(p_O + p_B)$, and 27.5 with likelihood $p_I(p_O + p_B)^2 + p_B(p_O + p_B)^2$. Gathering terms, we have

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

or, after factoring,

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

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

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

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

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

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

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

or

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

which is the condition in the text.

The remaining candidates for SNE are those in which players randomize over free-riding and one or more of the contribution strategies. To show that none of these are in fact viable, we note that attention can first be restricted to strategies of the form $\sigma^i = (p_F, 1 - p_F, 0, 0, 0)$: if there is some positive likelihood that each

of the others will free-ride, then profiles that sometimes call for the punishment of free-riders will fare worse than those that do not. The members of this restricted set can also be ruled out, however, since in the absence of punishment, contribution is dominated.

References

ACHESON, J. [1988], *The Lobster Gangs of Maine,* University Press of New England: Hanover, NH.

AXELROD, R. [1984], "An Evolutionary Approach to Norms," *American Political Science Review*, 80, 1095–1111.

BENDOR, J., AND D. MOOKHERJEE [1990], "Norms, Third-Party Sanctions and Cooperation," *The Journal of Law, Economics, & Organization*, 6, 33–63.

— AND P. SWISTAK [2001], "The Evolution of Norms," *American Journal of Sociology*, 106, 1493–1545.

BINMORE, K., J. GALE, AND L. SAMUELSON [1995], "Learning to Be Imperfect: The Ultimatum Game," *Games and Economic Behavior*, 8, 56–90.

— AND L. SAMUELSON [1994], "An Economist's Perspective on the Evolution of Norms," *Journal of Institutional and Theoretical Economics*, 150, 45–63.

— AND — [1999], "Evolutionary Drift and Equilibrium Selection," *The Review of Economic Studies*, 66, 363–393.

BOROFSKY, G., G. STOLLAK, AND L. MESSE [1971], "Sex Differences in Bystander Reactions to Physical Assault," *Journal of Experimental Social Psychology*, 7, 313–318.

BOWLES, S., AND H. GINTIS [2004], "The Evolution of Strong Reciprocity," *Theoretical Population Biology*, 65, 17–28.

BROOM, M., C. CANNINGS, AND G. VICKERS [1997], "Multi-Player Matrix Games," *Bulletin of Mathematical Biology*, 59, 931–952.

CARPENTER, J., S. BOWLES, H. GINTIS, AND S.-H. HWANG [2009], "Strong Reciprocity and Team Production," *Journal of Economic Behavior & Organization*, 71, 221–232.

— AND P. MATTHEWS [2005], "No Switchbacks: Rethinking Aspiration-Based Dynamics in the Miniature Ultimatum Game," *Theory and Decision*, 58, 351–385.

— AND — [2010], "Norm Enforcement: Anger, Indignation, or Reciprocity," *Journal of the European Economic Association*, forthcoming.

—, —, AND O. ONG'ONG'A [2004], "Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms," *Journal of Evolutionary Economics*, 14, 407–429.

DARLEY, J., AND C. D. BATSON [1973], "From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior," *Journal of Personality and Social Psychology*, 27, 100–108.

FEHR, E., AND U. FISCHBACHER [2004], "Third Party Punishment and Social Norms," *Evolution & Human Behavior*, 25, 63–87.

GINTIS, H. [2000], "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology*, 206, 169–179.

GÜTH, W., AND H. KLIEMT [1993], "Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation, and Moral Attitudes," *Metroeconomica*, 45, 155–187.

HUME, D. [1751/1983], *An Enquiry Concerning the Principles of Morals*, Hackett Publishing Company: Indianapolis.

LATANE, B., AND J. DARLEY [1970], *The Unresponsive Bystander: Why doesn't he Help?* Appleton-Century-Crofts: New York.

NACHBAR, J. H. [1990], "Evolutionary Selection Dynamics in Games: Convergence and Limit Properties," *International Journal of Game Theory*, 19, 59–89.

OLSON, M. [1965], *The Logic of Collective Action,* Harvard University Press: Cambridge, MA.

OSTROM, E. [1992], *Crafting Institutions for Self-Governing Irrigation Systems,* ICS Press: San Francisco, CA.

SCHLAG, K. [1998], "Why Imitate, and if so, How?" *Journal of Economic Theory*, 78, 130–156.

SETHI, R. [1996], "Evolutionary Stability and Social Norms," *Journal of Economic Behavior & Organization*, 29, 113–140.

SHOTLAND, L., AND M. STRAW [1976], "Bystander Response to an Assault: When a Man Attacks a Woman," *Journal of Personality and Social Psychology*, 34, 990–999.

TAYLOR, P., AND L. JONKER [1978], "Evolutionary Stable Strategies and Game Dynamics," *Mathematical Biosciences*, 40, 145–156.

Jeffrey P. Carpenter
Peter Hans Matthews
Department of Economics
Middlebury College
Middlebury, VT 05753
U.S.A.
E-mail:
jpc@middlebury.edu
pmatthew@middlebury.edu