

# AN EVOLUTIONARY MODEL OF RECIPROCITY

SUREN BASOV<sup>1</sup>

Department of Economics, The University of Melbourne, Melbourne,  
Victoria 3010, Australia.

---

<sup>1</sup>I am grateful to Peter Bardsley, Hsueh-Ling Huynh, Bob Rosenthal, Rajiv Sarin, Jay Surti, and Rabee Tourkey for helpful comments. The usual disclaimer applies.

**Abstract.** Despite the pervasiveness of reciprocal behavior, it has received little attention in the economic literature. In this paper, I consider an evolutionary model of reciprocity. The main findings of this paper are that evolution can support reciprocal behavior for the fraction of population, which is insensitive to the stakes involved, but is sensitive to the cohesiveness of the relationships. These findings match stylized facts learned from experimental and field studies of reciprocity.

# 1 INTRODUCTION

Despite the pervasiveness and potential economic importance of reciprocal behavior, it has received little attention in the economic literature. A reciprocal individual rewards nice behavior and punishes mean behavior, even at a personal cost. The last feature distinguishes reciprocity from altruism. Akerlof (1982) attempted to explain some features of wage setting in primary labor markets imposing a social norm of reciprocity, which called for an above-standard work performance in return for an above-market-clearing wage. However, the social norm was imposed exogenously rather than explained.

One way to explain cooperative behavior in general, and reciprocity in particular, is to argue that reciprocal behavior can be sustained as an equilibrium in a repeated game. This argument was developed, for example, in Baker, Gibbons, and Murphy (1997), Levine (1999).

However, there exists overwhelming experimental evidence provided, for example, by Fehr (2000) and Fehr and Gächter (2000), demonstrating that reciprocal behavior plays an important role even in the absence of repeated game effects.

Some stylized facts learned in experimental and field studies of reciprocity

are worth mentioning. First, the fraction of people who behave reciprocally is insensitive to the size of the stakes involved, but is strongly affected by the stability of the relationships in the group under study. Second, both self-interested and reciprocal individuals constitute a non-trivial fraction of the whole population. For a review of experimental studies, see Fehr and Gächter (2000). Results of the field studies are summarized in Mace (2000).

Another approach to explaining reciprocity is an evolutionary one. Some papers in this vein are Gintis (2000) and Sethi and Somanathan (2001a). For a review of evolutionary models, see Sethi and Somanathan (2001b). As argued by Sethi and Somanathan (2001b), four basic themes unite all evolutionary models of reciprocity, repetition, commitment, assortment, and parochialism. Repetition can give rise to evolution of the reciprocal behavior (Axelrod, 1984). In sporadic interactions, reciprocity is sustained either if selfish actions can be punished directly or indirectly, or if matching is sufficiently assortative.

However, in experimental studies (Fehr, 2000), even though neither of the four of above mentioned conditions held, a significant fraction of the experimental subjects exhibited reciprocal behavior. This finding motivates the point of view that reciprocity is an individual behavioral trait which is

allowed to change only rarely during life through a process of social learning. The main objective of this paper is to model this process and to study the dependence of the steady state fraction of reciprocal agents on the set of economic fundamentals. For this purpose, I develop an evolutionary model of reciprocity. I begin with a discrete time finite population model and then pass to a continuous limit. In this limit the evolution of the fraction of reciprocal agents will be governed by the replicator dynamics with aggregate shocks, similar to Foster and Young (1990).

There are several papers in the literature that propose behavioral models generating replicator dynamics in the limit. For examples, see Samuelson (1997) and Schlag (1998). They show that the replicator dynamics can be generated by a social learning procedure. Schlag (1999) also provides an example of a social learning procedure which leads to a general payoff monotone dynamic. However, it is a tricky business to write a behavioral model which converges to a replicator dynamics with aggregate noise. Models which introduce independent mutations (Kandori, Mailath, Rob (1993), Young 1993) will converge to a deterministically augmented replicator dynamics. To obtain a stochastic limit, aggregate shocks are needed. For an example of such a model, see Fudenberg and Harris (1992).

In my formulation of a behavioral model I follow closely Carradi and Sarin (2000). After constructing a model, I first study the deterministic limit and show that the replicator dynamics generates a continuum of equilibria; more precisely, if the share of reciprocal agents is below a certain cutoff, then no selective pressure operates and this share remains constant. However, there is another isolated asymptotically stable steady state that is characterized by a coexistence of positive shares of both reciprocal and purely self-interested individuals. The share of reciprocal agents in this steady state is determined by the set of economic fundamentals.

An interesting observation is that the equilibria with the share of reciprocal individuals below the above mentioned cutoff are unstable with respect to small aggregate shocks to utilities. Intuitively, since all agents get the same utility, there is no selective pressure in this region, and random changes to the share of reciprocal agents will accumulate. Once they pass the threshold (which will eventually happen with probability one), the replicator dynamics will take the system to the unique asymptotically stable equilibrium.

A flat-wage contract, which relies on reciprocal behavior, is an example of an underspecified contract. Hence, this paper can be viewed as a first step in providing evolutionary foundations for contractual underspecification.

Fehr and Gächter (2000) argued that contracts are often incomplete (or underspecified) in order to give agents an opportunity to reciprocate. While reciprocity is quite a common feature of human behavior, it is still not clear that a reciprocal contract will generate higher profits than the optimal incentive contract. For example, in the experiments conducted by Fehr (2000), though a majority of the subjects were reciprocal, the optimal incentive contract generated higher profits for the principal than the optimal reciprocal one. This might, however, change in the presence of bounded rationality or other factors that decrease the value of the second best incentive contract. In this case, reciprocity based contracts may become more attractive. This line of reasoning suggests that reciprocal contracts should be used more often to provide incentives for complex tasks than for simple ones.

## 2 THE MODEL

Consider a world that consists of  $2N$  workers and a finite number of firms. The firms are assumed to be profit maximizers. The workers can be of two types: self-interested or reciprocal. The employment history of a worker consists of a series of episodes of length  $\theta_N$ . The type of worker does

not change during the episode, but might change in the beginning of a new episode. Each episode consists of two periods.

The firms do not observe the type of each worker, but know the distribution of types in the population. They can offer two types of contracts: incentive contracts and trust contracts. Given an incentive contract, both types of workers react identically by choosing the optimal effort, which generates zero expected profits for the firms and zero expected utility for the workers. Given a trust contract, a self-interested worker shirks, generating expected utility  $U_2$  to herself and expected profits  $B$  to the firm; and a trustworthy worker exerts effort, generating expected utility  $U_1$  to herself and expected profits  $A$  to the firm. Assume that  $A > 0 > B$  and  $U_2 > U_1 > 0$ . Once all contracts are signed, the economy experiences an aggregate shock. After the end of the first period the worker may be fired. The probability of being fired is  $p_1^F$  if the worker did not shirk, and  $p_2^F$  if she did. Assume  $1 > p_2^F > p_1^F > 0$ , and that all firms observe whether the worker was fired in the first period of the current episode before offering the second period contract. Finally, define

$$\Delta U = (2 - p_1^F)U_1 - (2 - p_2^F)U_2,$$

and assume that  $0 < \Delta U < p_2^F U_2 - p_1^F U_1$ . At the end of an episode all the acquired information is destroyed.

At the beginning of a new episode there are  $N$  matchings with replacement in the workers' population. Only the first type drawn in a given pair is allowed to change her type. During any episode a learning opportunity arrives with probability  $\gamma\theta_N$ . The arrival of the learning opportunity is perfectly correlated among workers. If type  $j$  is matched with type  $k$  and the learning opportunity arrives, she observes the average payoffs  $v_j$  and  $v_k$  earned by types  $j$  and  $k$  in the episode and changes her type to  $k$  if and only if  $v_k > v_j$ .

**Assumption 1**  $v_i = u_i + \varepsilon_i$ , where  $u_i$  is the expected payoff of the type  $i$ <sup>2</sup>,  $\varepsilon_i$  is a random variable with zero mean.<sup>3</sup>

**Assumption 2**  $\varepsilon_j - \varepsilon_k$  is distributed uniformly on  $[-M, M]$ , where  $M > \sup_r |u_1 - u_2|$ .

Assumptions 1 and 2 imply

**Theorem 1** *If at least one worker of type  $j$  in  $jk$  pairs changes her type so do all other workers. The probability that all type  $j$  workers change their type*

---

<sup>2</sup> $u_i$  can depend on  $r$ .

<sup>3</sup>Under an aggregate shock different types might get different utilities because they might pursue different life styles.

is  $\theta_N q_{jk}$ , where  $q_{jk}$  is given by

$$q_{jk} = \gamma \left( \frac{1}{2} + \frac{u_k - u_j}{2M} \right). \quad (1)$$

Theorem 1 can be proven by a straightforward calculation and the proof is omitted. To proceed further, it is necessary to analyze the firms' behavior.

Let  $r$  be the proportion of reciprocal workers. The firms will offer a trust contract in period one if and only if

$$rA + (1 - r)B \geq 0. \quad (2)$$

Since workers and firms are a priori identical, all workers will receive the same contract in period one. If the firms offer an incentive contract in period one, they will offer incentive contracts in period two as well, since no new information is revealed. Suppose that the firms offered trust contracts in period one. In this case, their behavior in period two is described by the following:

**Theorem 2** *If the firms offered a trust contract in period one, they will offer a trust contract in period two to all workers if  $r > r^*$ , and a trust contract to the workers who were not fired and an incentive contract to the workers who*

were fired if  $r < r^*$ , where

$$r^* = \frac{|B|p_2^F}{Ap_1^F + |B|p_2^F}. \quad (3)$$

The proof of this theorem is in the Appendix.

If  $r = r^*$  the firms are indifferent between giving an incentive contract and a trust contract to the worker who was fired in period one. I will assume that in this case they mix with probability  $w$  given by

$$w = \frac{\Delta U}{p_2^F U_2 - p_1^F U_1}.$$

I want to study the evolution of the share of reciprocal agents in the limit of a large population and continuous time. The last limit is justified by the assumption that the length of an episode is small in comparison with the expected time between type changes, i. e. the type changes only rarely. To pass to this limit assume  $\theta_N = N^{-\delta}$ ,  $\delta \in (0, 1)$ . Restricting  $\delta$  to the open unit interval ensures that the length of an episode decreases with the size of population, and that within a unit of time a worker is matched with a vanishingly small share of population. Define  $\alpha = \gamma/M$ , then Proposition 2 of Corradi and Sarin (2000) implies that the share of reciprocal workers

follows a stochastic differential equation:

$$dr = \alpha\phi(r)dt + \gamma r(1-r)dZ. \quad (4)$$

Here  $Z$  is a standard Brownian motion and the function  $\phi(r)$  is defined as:

$$\phi(r) = \begin{cases} 0, & \text{if } r \in [0, \bar{r}] \cup \{r^*\}, \\ r(1-r)\Delta U, & \text{if } \bar{r} < r < r^*, \\ 2r(1-r)(U_1 - U_2), & \text{if } r^* < r \leq 1, \end{cases}$$

where  $\bar{r} = |B| / (A + |B|)$ .

To understand this equation, note that if  $r \leq \bar{r}$  the firms offer only incentive contracts, all workers get zero utilities, and there is no selective pressure, hence the share of reciprocal workers in the absence of shocks remains constant. If  $r^* < r \leq 1$ , all workers receive trust contracts in both periods, hence self-interested workers get higher utility and their share increases, while the share of reciprocal workers falls. In the region  $\bar{r} < r < r^*$  all workers get a trust contract in period one and a trust contract in period two if and only if they were not fired in period one. Under our assumptions on parameters of the model, reciprocal workers receive higher utility and their proportion increases.

Let us make a change of variables:  $\tau = \alpha t$  and  $\sigma^2 = \gamma^2/\alpha = \gamma M$ , and let  $W$  be a standard Brownian motion in time  $\tau$ . Then equation (4) takes the form

$$dr = \phi(r)d\tau + \sigma r(1-r)dW. \quad (5)$$

To obtain equation (5), I introduced a “slow” time in which learning opportunities arrive at a unit rate. Let us begin our analysis, considering the deterministic case  $\sigma \rightarrow 0$ . It is straightforward to observe that the steady states of the system are  $r \in [0, \bar{r}]$ ,  $r = 1$ , and  $r = r^*$ . The values  $r \in [0, \bar{r}]$  are Lyapunov (but not asymptotically) stable<sup>4</sup>, while  $r = 1$  and  $r = \bar{r}$  are unstable. The value  $r = r^*$  is asymptotically stable. This implies that if the initial share of reciprocal agents  $r_0 > \bar{r}$ , then in the long run it will tend to  $r^*$ .

The deterministic system possesses a continuum of steady states. Hence, an equilibrium selection problem arises. To select one equilibrium, I will use the concept of stochastic stability (Friedlin and Wentzel, 1984). For this purpose I prove the following theorem:

---

<sup>4</sup>A stationary point  $r_0$  of an ordinary is called Lyapunov stable if  $\forall \varepsilon > 0 \exists \delta > 0$  such that if the initial position of the system is within  $\delta$  of  $r_0$  it will remain within  $\varepsilon$  forever. If the system is Lyapunov stable and converges to  $r_0$  as time goes to infinity, provided the initial position was within  $\delta$  from  $r_0$ , it is called asymptotically stable.

**Theorem 3** *Suppose that evolution of the share of the reciprocal workers in the population is governed by equation (5), and that  $r(0) \in (0, 1)$ . Let  $F(r, t, \sigma)$  denote the probability that at time  $t$  this share is below  $r$ . Then there exists a unique asymptotically stable stationary distribution  $F(r, \sigma)$  that is consistent with the equation (5) and*

$$\lim_{\sigma \rightarrow 0} F(r, \sigma) = 1 \text{ for } r \geq r^*$$

$$\lim_{\sigma \rightarrow 0} F(r, \sigma) = 0 \text{ for } r < r^*.$$

The proof is in the Appendix. The theorem implies that if the workers change their type rarely, the share of the reciprocal workers stays arbitrary close to  $r^*$  at almost all the time.

### 3 DISCUSSION AND CONCLUSIONS

In this paper I demonstrated that evolution can support reciprocal behavior in a fraction of the population. The model developed in this paper differs from most papers in the literature, since it does not rely on repetition, assortment, commitment, and parochialism. Reciprocity is considered rather as a simple behavioral trait. Due to this fact, this model better fits a typical experimental setup.

The stochastically stable fraction of reciprocal individuals in this model is always strictly between zero and one. In Fehr's (2000) experiments, in the trust treatment approximately 30% of the subjects behaved in a self-interested way; i. e., exerted the lowest possible effort for a fixed wage. Hence, the fraction of reciprocal agents was approximately 70%. It is also interesting to note that  $r^*$  and  $\bar{r}$  are invariant with respect to rescaling the payoffs. They depend, however, on the probabilities of separation, in particular on  $p_1^F$ , the probability a worker who is not guilty of cheating is fired, which can be viewed as a measure of the cohesiveness of the firm-worker relationship. This matches the stylized facts reported in Mace (2000).

Some comparative statics results are worth mentioning. First, note that as  $p_1^F \rightarrow 0$ , the share of reciprocal types  $r^* \rightarrow 1$ . Hence, if reciprocal workers do not get fired the entire population becomes asymptotically reciprocal. On the other hand, as  $p_1^F \rightarrow 1$  reciprocity disappears, since  $r^* \rightarrow \bar{r}$ , the boundary of the incentive contracts region. This suggests that in stable economies with low rates of job separation, reciprocity will be a more common phenomenon than in economies where job separation occurs more often. This prediction seems similar to the one that can be generated through a repeated game mechanism. However, there is one significant difference. A repeated

game model will predict that workers hired by the firm for life will exhibit reciprocal behavior towards the firm, but they need not exhibit such behavior outside this particular relationship. The model of this paper, on the contrary, predicts that workers from economies with stable employment will exhibit reciprocal behavior in all their relations.

## BIBLIOGRAPHY

- Akerlof, G. A. (1982). "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, 97, pp.543-569.
- Anderson S. P., J. K. Goeree, and C. A. Holt. (1997) *Stochastic Game Theory: Adjustment to Equilibrium under Bounded Rationality*, unpublished draft.
- Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.
- Baker, G., R. Gibbons, and K. J. Murphy. (1997) *Relational Contracts and Theory of the Firm*, mimeo, MIT.
- Basov, S. (2001). "Bounded Rationality, Reciprocity, and Their Economic Consequences," Ph.D. Dissertation, Graduate School of Arts and Sciences, Boston University.
- Corradi, V., and R. Sarin (2000). "Continuous Approximation of Stochastic Evolutionary Game Dynamics," *Journal of Economic Theory*, 94, pp.163-191.
- Fehr, E., and S. Gächter. (2000). *Fairness and Retaliation: The Economics of Reciprocity*, Working Paper No. 40, University of Zurich.
- Fehr, E. (2000). *Do Incentive Contracts Crowd Out Voluntary Cooperation?* mimeo, University of Zurich.

- Foster, D., and P. Young (1990). "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, 38, pp.219-232.
- Freidlin, M., and Wentzel. (1984). *Random Perturbations of Dynamical Systems*. New York: Springer.
- Fudenberg, D., and C. Harris. (1992). "Evolutionary Dynamics with Aggregate Shocks," *Journal of Economic Theory*, 57, pp. 420-441.
- Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology*, 205: 1-11.
- Kandori, M., G. Mailath, and R. Rob. (1993). "Learning, Mutation and Long Run Equilibria in Games," *Econometrica*, 61, pp. 29-56.
- Levin, J. (1999). *Relational Incentive Contracts*, mimeo, MIT.
- Mace, R. (2000). Human behavior: Fair game. *Nature*, 406: 248-49.
- Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection*, MIT Press.
- Schlag, K. H. (1999). Which One Should I Imitate? *Journal of Mathematical Economics*, 31: 493-522.
- Schlag, K. H. (1998). Why Imitate and if So, How? A Boundedly Rational Approach to Multiarmed Bandits. *Journal of Economic Theory*, 78: 130-56.
- Sethi, R., and E. Somanathan. (2001a). "Preference Evolution and Reci-

procity,” *Journal of Economic Theory*, 97, 273-297.

Sethi, R., and E. Somanathan. (2001b). “Understanding Reciprocity,” *Journal of Economic Behavior and Organization*, forthcoming.

Young, P. (1993). “The Evolution of Conventions,” *Econometrica*, 61, pp. 57-84.

## APPENDIX

In this appendix I will give proofs of Theorem 2 and 3.

Theorem 2:

**Proof.** The probability that the worker is reciprocal conditional on the event that she was not fired is given by:

$$p(R \mid \text{not fired}) = \frac{(1 - p_1^F)r}{(1 - p_1^F)r + (1 - p_2^F)(1 - r)} > r. \quad (6)$$

Hence, since the firms found it in their interest to offer a trust contract in period one, they will a fortiori find it in their interest in period two after observing the outcome “not fired.”

The probability that the worker is reciprocal conditional on the event that she was fired is given by

$$p(R \mid \text{fired}) = \frac{p_1^F r}{p_1^F r + p_2^F (1 - r)}. \quad (7)$$

The expected payoff to the firm from a trust contract with a worker who was fired is:

$$\frac{Ap_1^F r + Bp_2^F(1-r)}{(1-p_1^F)r + (1-p_2^F)(1-r)}, \quad (8)$$

which is nonnegative if and only if  $r \geq r^*$ . ■

Theorem 3:

**Proof.** Equation (5) implies that  $F(r, \sigma)$  is governed by the following partial differential equation:

$$\phi(r) \frac{\partial F}{\partial r} = \frac{\sigma^2}{2} \frac{\partial}{\partial r} \left( r(1-r) \frac{\partial F}{\partial r} \right), \quad (9)$$

subject to the boundary conditions  $F(0, t) = 0$  and  $F(1, t) = 1$ . Introduce the function

$$\Pi(r) = \int_0^r \frac{\phi(x) - \frac{\sigma^2}{2} x(1-x)}{x(1-x)} dx. \quad (10)$$

Then the unique asymptotically stable solution is given by

$$F(r, \sigma) = \frac{\int_0^r \exp\left(\frac{2\Pi(x)}{\sigma^2}\right) dx}{\int_0^1 \exp\left(\frac{2\Pi(x)}{\sigma^2}\right) dx}. \quad (11)$$

(For a derivation of equation (9) and discussion of its properties, see Anderson, Goeree, Holt (1997)). For a sufficiently small  $\sigma$  the function  $\Pi(r)$

decreases for  $r > r^*$  and increases for  $r < r^*$ , hence, it achieves its maximum at  $r = r^*$ . Write (11) in a form

$$F(r, \sigma) = \frac{\int_0^r \exp\left(\frac{2(\Pi(x) - \Pi(r^*))}{\sigma^2}\right) dx}{\int_0^1 \exp\left(\frac{2(\Pi(x) - \Pi(r^*))}{\sigma^2}\right) dx}. \quad (12)$$

Let  $V(x) = \Pi(r^*) - \Pi(x)$ . Note that  $V(x) \geq 0$  for any  $x \in (0, 1)$ , and  $V(x) = 0$  if and only if  $x = r^*$ . For any Borel  $D \subset R$  define

$$\mu_\sigma(D) = \int_D dF(x, \sigma).$$

Then (see Friedlin and Wentzel, 1984)

$$\sigma^2 \ln \mu_\sigma(D) \rightarrow \inf_{x \in D} V(x).$$

Let  $D$  be a closed set such that  $r^* \notin D$ . Then  $\inf_{x \in D} V(x) < 0$ . Hence  $\mu_\sigma(D) \rightarrow 0$  and  $\mu_\sigma(R/D) \rightarrow 1$  when  $\sigma$  approaches zero. Let  $D'$  be an open set such that  $r^* \in D'$ . Then the set  $R/D'$  is closed and  $r^* \notin R/D'$ . Hence,  $\mu_\sigma(R/D) \rightarrow 0$  and  $\mu_\sigma(D') \rightarrow 1$  as  $\sigma$  approaches zero. But this together with the left continuity of  $F(r, \sigma)$  implies that  $F(r, \sigma)$  has the required form. ■