

Crying Over Spilt Milk: Sunk Costs, Fairness Norms and the Hold-Up Problem*

by

Jeffrey P. Carpenter and Peter Hans Matthews

June 2003

MIDDLEBURY COLLEGE ECONOMICS DISCUSSION PAPER NO. 03-12



DEPARTMENT OF ECONOMICS
MIDDLEBURY COLLEGE
MIDDLEBURY, VERMONT 05753

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

Crying Over Spilt Milk: Sunk Costs, Fairness Norms and the Hold-Up Problem*

Jeffrey P. Carpenter
Department of Economics
Middlebury College
Middlebury, Vermont 05753
jpc@middlebury.edu

Peter Hans Matthews
Department of Economics
Middlebury College
Middlebury, Vermont 05753
pmatthew@middlebury.edu

June 12, 2003

Abstract

This paper explores a possible connection between two behavioral anomalies in economics, the observed responsiveness of individual decision-makers to sunk costs, and the apparent failure of backward induction to predict outcomes in experimental bargaining games. In particular, we show that under some conditions, a "sunk cost sensitive" fairness norm can evolve in such environments. Under this norm, a fair distribution allows all parties to recoup whatever each has invested in their relationship before the *net* surplus is then divided into equal shares. The establishment of such a norm would have important consequences for the hold-up problem, which we characterize in terms of ultimatum bargaining in the presence of an outside option. We then conclude with a brief discussion of the possible labor market implications of our results.

JEL Classification: C78, J41

Keywords: sunk costs, norms, fairness, trust, hold-up problem, human capital

*We thank Stephen Burks, Carolyn Craven and Corinna Noellke for their comments. The first author also thanks the National Science Foundation (SES-CAREER 0092953) for financial support.

1 Introduction

Our purpose in this brief paper is to explore a possible connection between two important behavioral anomalies in economics. The first of these is the observed responsiveness of individual decision-makers to sunk costs, both their own and, in some cases, those of others.¹ Textbooks often present the "irrelevance of sunk costs" as an almost canonical principle: Mankiw [1998, 291], for example, describes it a "deep truth about rational decision-making," a codification of the adage "don't cry over spilt milk." For more than two decades, however, at least since the publication of Thaler [1980], behavioral psychologists have documented the effects of "spilt milk" on human behavior. Experimental economists have contributed to this literature as well: Phillips, Battalio and Kogut [1991] and Hackett [1993], for example, examine the influence of sunk costs on, respectively, individual valuation and bargaining. Given considerable evidence that spilt milk sometimes matters to reasonable people, economists should be reluctant to dismiss such behavior as *prima facie* "irrational."

The second of these anomalies is the limited predictive power of backward induction in even simple bargaining games. Güth, Schmittberger and Schwarze's [1982] experiments were perhaps the first to demonstrate that "selfish" offers were less common, and rejected more often, in the ultimatum game than was consistent with subgame perfection, a result replicated dozens of times since.

One of the most influential explanations of the Güth *et al* results involves the evolution of a robust fairness norm. Consider the miniature ultimatum game or MUG described in Binmore, Gale and Samuelson [1995], a reference point for much of our subsequent discussion. Two players, A and B , must divide a surplus of 4 units. B first proposes one of two allocations, a "fair" offer in which A and B each receive 2, and a "selfish" one, in which B receives 3 and A receives 1. A can either accept fair offers and reject selfish ones, where rejection leaves both A and B with 0, or accept all offers, fair or selfish. In the subgame perfect equilibrium or SPE, B 's proposal is selfish but A accepts it. Even in this transparent environment, however, the experimental evidence is more consistent with the second, non-SPE, Nash equilibrium of MUG, in which B proposes a fair division and A randomizes such that selfish offers are rejected at least one third of the time.²

Binmore et al showed, however, that if MUG was (re)framed as an evolutionary game, all of the population states corresponding to the continuum of Nash equilibria (except the endpoint) were neutrally stable under the so-called replicator dynamic. (On the other hand, the single state that corresponds to the SPE was asymptotically stable.) Furthermore, for some forms of drift, this continuum collapsed into an asymptotically stable state in which almost A s proposed fair allocations. If and when this occurs, evolution toward this state has

¹The common belief that merchants should not price gouge after a natural disaster is an example of the latter.

²For a recent attempt to achieve an even closer reconciliation of model and data, see Carpenter and Matthews [2003].

sometimes been characterized in terms of the establishment of a robust fairness norm.

What this model does not explain is the demarcation between "fair" and "unfair." Since the surplus in MUG materializes out of thin air - that is, no investment is required to produce it - the fair distribution leaves A and B with equal shares of both the *gross* and *net* surplus. If some investment is required, however, these two notions of fairness must be differentiated. In the next three sections, we consider if and when such norms could evolve, first in isolation and then in combination. The evolution of such norms have important implications for the hold-up problem (Williamson 1975), and this will frame much of our discussion. We then conclude with a brief discussion of the implications of our model for hold-up in labor markets and possible further research.

Our work is also related to the recent contributions of Carmichael and MacLeod [2002] and Ellingsen and Robles [2002] to the hold-up literature. The former show that "sunk costs and an equal share of the net surplus" is a strict Nash equilibrium, and therefore evolutionarily stable, in a modified Groot [1984] model with continuous investment levels, in which bargaining assumes the form of the Nash demand game. In contrast, the latter conclude that "stochastic stability has no cutting power" when generalized ultimatum bargaining follows the choice of continuous investment levels. Neither paper considers the selection mechanism, or possible mutation/drift, in much detail, however.

2 The Strategic Environment

We consider three variations of MUG with an outside option, perhaps the simplest representation of the hold-up problem. In all three, A must first decide whether or not to invest in some relationship with B , at a cost c to herself, where $1 < c < 2$. If she does not invest, then both A and B receive 0; if she does, a post-investment surplus of 4 units is produced, which B must then propose how to allocate. In all variations, a fixed "selfish" offer, in which B reserves 3 units for himself, is available. The differences are found in the "fair" offers that are also possible. In the first variation, which we denote HUG1, B can also propose a "weakly fair" distribution, in which A and B would receive equal shares of the gross surplus, and which leaves A with $2 - c$ after the costs of investment. In the second, HUG2, the alternative to selfishness is "strong fairness," under which A is first allowed to recoup these costs before the net surplus is shared, a distribution that leaves both A and B with $2 - (c/2)$. In the third and final variation, HUG3, B can be selfish or fair in either sense. In all cases, A must then decide whether or not to accept B 's proposal: if she turns it down, her investment is lost, but B receives 0. To provide a more evocative frame for our models, we shall call A 's options "accept selfish offers or better," "accept (only) weakly fair offers or better" and "accept (only) strongly fair offers of better," and observe that the third embodies a "sunk cost sensitive" notion of fairness,

one that varies with the level of investment c . The normal forms for HUG1, HUG2 and HUG3 are then as depicted in Figures 1 and 2.

[Insert Figures 1 and 2 About Here]

Within this particular framework, there is a "hold-up problem" because in all three variations, the SPE is one in which A does not invest, a decision that leaves both A and B worse off than if A had invested, B had responded with either sort of fair offer, and A had accepted it. The recent experimental work of Oosterbeek *et al* (1998) tells us, however, that SPE's predictive power is no better, indeed worse, in this sort of framework. It is therefore important to note that all three also have components of non-SPE Nash equilibria in which A does invest, B 's offer is fair and A is sometimes prepared to turn down selfish offers. If either sort of norm is to have a role, then this outcome, which seems to be predicated on an incredible threat to abandon the post-investment relationship if the proposal is lopsided, must be rationalized.

To this end, we consider an environment in which the A s and B s are drawn from two large, distinct populations of equal size, in which the A s are matched at random to the B s at discrete intervals Δ . As a consequence, perhaps, of the rules or norms that inform individual behavior in uncertain environments, we suppose that members of both populations are "predisposed" to use one of the pure strategies available to them, but that these predispositions are not immutable. In particular, a small fraction θ^A of the A s (and θ^B of the B s) are assumed to "drop out" of **HUG** between rounds, and replaced with new members whose initial norms/behaviors are to some extent random. (One could also assume, in the spirit of Binmore et al, that θ^A and θ^B are rates of decision error, about which more below.) In addition, of the $1 - \theta^A$ A s, and $1 - \theta^B$ B s that remain, a proportion $\mu = z\Delta$, z fixed, are assumed to (re)evaluate their respective situations between rounds. Consistent with Schlag (1994), each of these A s (B s) samples another A (B) at random, learns his or her behavior and outcome, and then imitates him or her if the difference is positive and exceeds some switching cost q , the value of which is drawn from a uniform *pdf* $[0, \tilde{q}]$. We assume that $\tilde{q} \geq 3$, which ensures that the likelihood of a switch lies between 0 and 1, and observe that since the mean switch cost is $\tilde{q}/2$, \tilde{q} can be interpreted as a measure of "norm adherence."

To formalize the laws of motion consistent with this behavior, some final notation is needed. Let $p_{SF}^A(t)$ be the proportion of A s in HUG2 or HUG3 who invest and accept (only) strongly fair offers or better in round t ; $p_{WF}^A(t)$, the proportion of A s in HUG1 or HUG3 who invest and accept (only) weakly fair offers or better; $p_S^A(t)$, the proportion of A s in HUG1, HUG2 or HUG3 who invest and accept selfish or better (that is, all) offers; and $p_D^A(t)$, the proportion of A s in HUG1, HUG2 or HUG3 who do not invest. Likewise, let $p_{SF}^B(t)$ be the proportion of B s in HUG2 or HUG3 who propose strongly fair allocations; $p_{WF}^B(t)$, the proportion of B s in HUG1 or HUG3 who propose weakly fair allocations; and last, $p_S^B(t)$, the proportion of B s in HUG1, HUG2 or HUG3 whose proposals are selfish.

The evolution of population shares will then follow:

$$p_j^i(t + \Delta) - p_j^i(t) = (1 - \theta^i)z\Delta\tilde{q}^{-1}p_j^i(t)\{\sum_{k \neq j} p_k^i(t)[\pi_j^i(t) - \pi_k^i(t)]^+ (2.1) \\ - \sum_{k \neq j} p_k^i(t)[\pi_k^i(t) - \pi_j^i(t), 0]^+\} + \theta^i(d_j^i - p_j^i(t))$$

where $[x]^+ = \max[x, 0]$, d_j^i is the share of "newcomers" to i who adopt j , where $\sum_j d_j^i = 1$, and $\pi_j^i(t)$ is the expected payoff to the members of i committed to j in round t . It is not difficult to show that (2.1) can be rewritten:

$$\frac{p_j^i(t + \Delta) - p_j^i(t)}{\Delta} = (1 - \theta^i)z\tilde{q}^{-1}p_j^i(t)(\pi_j^i(t) - \bar{\pi}^i(t)) + \theta^i(d_j^i - p_j^i(t)) \quad (2.2)$$

where $\bar{\pi}^i(t)$ is the population-wide mean payoff. As $\Delta \rightarrow 0$, (2.2) becomes a continuous time process with both "selection" and "drift":

$$\dot{p}_j^i = (1 - \theta^i)z\tilde{q}^{-1}p_j^i(\pi_j^i - \bar{\pi}^i) + \theta^i(d_j^i - p_j^i) \quad (2.3)$$

where dots denote time derivatives and the dependence of p_j^i , π_j^i and $\bar{\pi}^i$ on t has been suppressed. The selection mechanism assumes the familiar form of a (scaled, in this case) replicator dynamic. In the absence of drift, when $\theta^A = \theta^B = 0$, neither the amount of norm adherence \tilde{q} nor the relative speed of self-evaluation affects the evolution of shares, but once "newcomers" are introduced, both then matter.

Consider the behavior of population shares in HUG1, for example. Since the replicator dynamic for each population is domain invariant, two "residual shares" - in this case, the proportion of A s who do not invest p_D^A and the proportion of B s who are selfish p_S^B - can be eliminated, and attention restricted to a three dimensional process. Substitution for $p_D^A = 1 - p_{WF}^A - p_S^A$ and $p_S^B = 1 - p_{WF}^B$ and some simplification leads to:

$$\begin{aligned} \dot{p}_{WF}^A &= (1 - \theta^A)z\tilde{q}^{-1}p_{WF}^A[(1 - p_{WF}^A)(2p_{WF}^B - c) - p_S^A(p_{WF}^B + 1 - c)] \\ &\quad + \theta^A(d_{WF}^A - p_{WF}^A) \\ \dot{p}_S^A &= (1 - \theta^A)z\tilde{q}^{-1}p_S^A[(1 - p_{WF}^A)(p_{WF}^B + 1 - c) - p_{WF}^A(2p_{WF}^B - c)] \\ &\quad + \theta^A(d_S^A - p_S^A) \\ \dot{p}_{WF}^B &= (1 - \theta^B)z\tilde{q}^{-1}p_{WF}^B(1 - p_{WF}^B)(2p_{WF}^A - p_S^A) + \theta^B(d_{WF}^B - p_{WF}^B) \end{aligned} \quad (2.4)$$

In a similar vein, the evolution of shares in HUG2 will follow:

$$\begin{aligned}
\dot{p}_{SF}^A &= (1 - \theta^A)z\tilde{q}^{-1}p_{SF}^A[(1 - p_{SF}^A - p_S^A)p_{SF}^B(2 - (c/2)) \\
&\quad - (1 - p_{SF}^B)(c(1 - p_{SF}^A - p_S^A) - p_S^A)] + \theta^A(d_{WF}^B - p_{SF}^A) \\
\dot{p}_S^A &= (1 - \theta^A)z\tilde{q}^{-1}p_S^A[p_{SF}^B(1 - p_{SF}^A - p_S^A)(2 - (c/2)) \\
&\quad + (1 - p_{SF}^B)((1 - p_S^A) - c(1 - p_{SF}^A - p_S^A))] + \theta^A(d_S^A - p_S^A) \\
\dot{p}_{SF}^B &= (1 - \theta^B)z\tilde{q}^{-1}p_{SF}^B(1 - p_{SF}^B)[(p_{SF}^A + p_S^A)(2 - (c/2)) - 3p_S^A] \\
&\quad + \theta^B(d_{WF}^B - p_{SF}^B)
\end{aligned} \tag{2.5}$$

Because there are more options for both A s and B s in HUG3, however, its laws of motion are somewhat more complicated:

$$\begin{aligned}
\dot{p}_{SF}^A &= (1 - \theta^A)z\tilde{q}^{-1}p_{SF}^A[(1 - \Sigma^A)(p_{SF}^B(2 + (c/2)) - c) - 2(p_{WF}^A + p_S^A)p_{SF}^B \\
&\quad - p_S^A(1 - \Sigma^B)] + \theta^A(d_{SF}^A - p_{SF}^A) \\
\dot{p}_{WF}^A &= (1 - \theta^A)z\tilde{q}^{-1}p_{WF}^A[(1 - \Sigma^A)(p_{SF}^B(2 + (c/2)) - c) + 2(1 - p_{WF}^A - p_S^A)p_{WF}^B \\
&\quad - p_S^A(1 - \Sigma^B)] + \theta^A(d_{WF}^A - p_{WF}^A) \\
\dot{p}_S^A &= (1 - \theta^A)z\tilde{q}^{-1}p_S^A[(1 - \Sigma^A)(p_{SF}^B(2 + (c/2)) - c) + 2(1 - p_{WF}^A - p_S^A)p_{WF}^B \\
&\quad + (1 - p_S^A)(1 - \Sigma^B)] + \theta^A(d_S^A - p_S^A) \\
\dot{p}_{SF}^B &= (1 - \theta^B)z\tilde{q}^{-1}p_{SF}^B[(1 - p_{SF}^B)(\Sigma^A(2 + (c/2)) - 3p_S^A) - p_{WF}^B(2p_{SF}^A - p_S^A)] \\
&\quad + \theta^B(d_{SF}^B - p_{SF}^B) \\
\dot{p}_{WF}^B &= (1 - \theta^B)z\tilde{q}^{-1}p_{WF}^B[(1 - p_{WF}^B)(2p_{SF}^A - p_S^A) - p_{SF}^B(\Sigma^A(2 + (c/2)) - 3p_S^A)] \\
&\quad + \theta^B(d_{WF}^B - p_{WF}^B)
\end{aligned} \tag{2.6}$$

where $\Sigma^A = p_{SF}^A + p_{WF}^A + p_S^A$ and $\Sigma^B = p_{SF}^B + p_{WF}^B$.

3 The Fairness Norms in Isolation

We first consider the fortunes of the weak fairness norm in isolation. As noted in the previous section, HUG1 has two components of Nash equilibria. In the first, denoted C_1^1 , A does not invest and B randomizes between selfishness and weak fairness, such that the former is chosen at least $c - 1$ percent of the time. This component includes the SPE, and as the cost of investment c tends toward 2, B never "experiments" with fairness. In the second, C_2^1 , A does invest and is prepared to turn down a selfish offer at least one third of the time, and B 's proposal is (always) weakly fair.

The "pseudo phase diagrams" in Figures 3a and 3b depict the evolution of population shares in HUG1 in the absence of drift when the cost of investment is low $c = 1.25$ and high $(c = 1.75)$. On the horizontal axis, we measure the sum of shares $p_{WF}^A + p_S^A = 1 - p_D^A$ or, in words, the proportion of A s who invest, norm-driven or not, and on the vertical, the proportion of B s who propose weakly fair

distributions p_{WF}^B . Since each point is consistent with various combinations of p_{WF}^A and p_S^A , the solution paths can and do cross - it is sometimes helpful to interpret the picture as a projection of sorts - and the component of particular interest, C_2^1 , "shrinks" to the (1, 1) vertex. To minimize the loss of information, we choose initial conditions so that each of the values of $p_{WF}^B(0)$ selected is matched with three (constant sum) pairs of $p_{WF}^A(0)$ and $p_S^A(0)$, one balanced and two unbalanced. (For example, the three solution paths that emanate from the point (0.25, 0.25) have the initial values ($p_{WF}^A(0) = 0.125, p_S^A(0) = 0.125, p_{WF}^B(0) = 0.25$), ($p_{WF}^A(0) = 0.1875, p_S^A(0) = 0.0625, p_{WF}^B(0) = 0.25$) and ($p_{WF}^A(0) = 0.0625, p_S^A(0) = 0.1875, p_{WF}^B(0) = 0.25$).

[Insert Figures 3a Through 3h About Here]

What these phase plots show is that both components are (at least weak) attractors, and that the basin of attraction for the "no hold-up equilibrium" becomes smaller as investment costs rise. The latter is consistent with the intuition that as the stakes rise, the likelihood of a norm-based solution to the hold-up problem tends to fall. Absent turnover, decision error or other unmodelled "mutation," then, there is reason to believe that the As and Bs will *sometimes* reach the efficient outcome on their own. The problem, however, is that because no element of (in particular) the second component is isolated, one cannot conclude *a priori* that it is "drift compatible" (Binmore and Samuelson 1999).

The introduction of drift requires a full(er) parametrization of the model. Our initial choices for the selection mechanism in HUG1 are $z = 1$ and $\tilde{q} = 5$ and for the drift function, $\theta^A = \theta^B = 0.01$, consistent with a turnover rate of one percent in both populations, and $d_{WF}^A = d_S^A (= d_D^A) = 1/3$ and $d_{WF}^B (= d_S^B) = 1/2$, which we shall refer to as "neutral drift." (It is neutral in the sense that newcomers are not predisposed to adopt one behavioral rule over another.) Figures 3c and 3d, constructed on the same lines as the first of these diagrams, depict the evolution of population shares under these conditions.

The most important feature of this evolution, the existence of a unique stable equilibrium "near" the state that corresponds to the SPE, will not come as much surprise to those familiar with Binmore et al's simulation of behavior in MUG: in intuitive terms, the number of new As who are endowed with the fairness norm is too small, relative to the number of Bs who are fair, to stabilize C_2^1 . The surprise, perhaps, is how far "near" can be. Table 1, which records all of the stable rest points for HUG1 and HUG2 under various scenarios, shows that the proportion of As who do not invest when $c = 1.25$ is "just" 82.3%, and a little more than 10% of those who do not (1.8% of the total) would still turn down an unfair offer, while 16.5% of the Bs propose equitable distributions. As investment costs rise, however, so does the proportion of As who do not invest, from 82.3% to 93.1%, as does the share of the much smaller number who do invest and would turn down a selfish offer, to 43.6%. There is an important implication for experimental research here: even in situations where participants do not, or perhaps cannot, "solve the hold-up problem," and even

if the predominant fairness norm is weak, the number of investors, however small, and the number of those who propose a fair allocation, will be sunk cost sensitive.

[Insert Table 1 About Here]

The second component is also drift compatible, however. To see this, consider the evolution of population shares when drift is still neutral and θ^B remains fixed at 0.01, but θ^A is increased, to 0.075. Under our interpretation of the θ s as turnover rates, if the *As* and *Bs* are identified as, respectively, workers and firms, then this is consistent with the reasonable view that workers enter and exit labor markets more often than firms. If, in the spirit of Binmore et al, these are treated as probabilities of decision error, this is plausible if, for population states close to the second component, when the difference between π_{WF}^A and π_S^A is small, the members of *A* become more prone to mistakes. There will now be *two* stable equilibria, as illustrated in Figures 3e and 3f and reported in Table 1. Most important, in one of these, the *As* and *Bs* have resolved the hold-up problem: when investment costs are low, fewer than 10% of all *As* do not invest and about 90% of the *Bs* propose fair allocations of the surplus. Furthermore, of the more than 90% of *As* who do invest, 40.6% (or 36.8% of the total) will be norm-driven and so turn down the infrequent selfish offer. As investment costs rise, there is substantial increase (to almost 30%) in the number of *As* who don't invest in the no hold-up equilibrium, but almost no effect on the behavior of the *Bs*. If the weak fairness norm "takes root," in other words, both the decision to invest, and the behavior of those who have invested, will seem to be functions of the costs of investment, even if the norm is itself not "sunk cost sensitive," results that are, in principle, testable in the experimental lab.

But will such sunk cost sensitive norms ever take root themselves? Like HUG1, HUG2 has two Nash components, but in this case, *both* are dependent on c . In the first, C_1^2 , *A* does not invest and *B* randomizes, such that the proposal is selfish at least $2(c-1)/(2+c)$ percent of the time, and in the second, C_2^2 , *A* invests and randomizes such that selfish offers would be turned at least $(2+c)/6$ percent of the time, and *B*'s proposed distribution is always strongly fair. As illustrated in Figures 4a and 4b, both C_1^2 and C_2^2 are (at least) neutrally stable if there is no drift.³ Consistent with intuition, both the size of the second, no hold-up, component, as well as its basin of attraction, decrease in size as the cost of investment rises.

[Insert Figures 4a Through 4h About Here]

There is reason to believe, however, that drift will be "less kind" to those with sunk cost sensitive norms: confronted with fixed proportions of *As* who do invest, invest and accept selfish offers and invest but insist on a (weak in HUG1, strong in HUG2) fair distribution, *Bs* will find it more expensive to accommodate

³The pseudo phase plots for HUG2 are constructed on the same lines as those for HUG1, except that the proportion of *As* who invest, measured on the horizontal axis, is now $p_{SF}^A + p_S^A$, and the proportion of *Bs* who are fair, on the vertical axis, is p_{SF}^B .

such norms. It comes as little surprise, then, that in the benchmark case where turnover rates are small (one percent) and equal, the one stable equilibrium is that which corresponds to the SPE, as depicted in Figures 4c and 4d. As before, however, the surprise is that when investment costs are low, almost one fifth (17.0%) of all *As* do invest, and close to 10% of all the offers are strongly fair. But as investment costs rise, the proportion of *As* who invest falls, to 6.6%, but the proportion of those who invest and receive fair offers rises, to more than 20%. The implications for experimentalists and other empirical researchers are once more clear: even if "players" fail to solve the hold-up problem on their own, their behavior in the presence of even small "noise" will seem to depend on sunk costs.

The intuition that it is more difficult for sunk cost sensitive norms to establish themselves is confirmed in Figures 4e and 4f, which illustrate the evolution of shares when the turnover rate for *A* is increased to 7.5%. This was more than sufficient to stabilize the no hold-up component in HUG1 (see above) but fails to do so in HUG2. The "almost SPE" state is once more the unique stable equilibrium. It is important to note, however, that when investment costs are low, more than 40% (!) of the *As* invest, even if less than 6% receive fair offers. As investment costs rise, the former share falls, to 25.9%, while the latter rises, to almost 10%, the same pattern observed in the case of equal turnover rates.

This prompts the question: is the hold-up problem *ever* resolved in HUG2? Given the simple form that drift assumes in our model, and the choice of parametrization, not least the one percent turnover rate for the *Bs*, there is no *plausible* turnover rate for the *As* that does so. It does not follow, however, that the second component is "drift incompatible": Binmore and Samuelson's [1999] demonstration that "strict path Nash equilibria" are drift compatible assumes that drift rates are outcome sensitive. In fact, if we set $\theta^A = 0.075$ and $\theta^B = 0.01$ and further assume that drift is *prosocial* - in particular, $d_{SF}^A = (2/3)$ and $d_S^A = d_D^A = (1/6)$, which implies that two thirds of the new *As* adopt the fairness norm - then there will be two stable equilibria, as shown in Figures 4g and 4h. One of these is still a low investment state, but in the other, more than 95% of all the *As* invest, almost 80% of these are prepared to "punish" a selfish offer, and more than 95% of all the offers are fair when sunk costs are small. Furthermore, as costs rise, the proportion of those who invest falls (just) a little, as does the proportion of fair offers. Once the *As* and *Bs* have resolved the no hold-up problem, it will seem, from the perspective of the experimenter, that the *As* will become more reluctant to invest, and the *Bs* more opportunistic, as the stakes rise.

Last, we note as an aside that when this "unequal and prosocial" drift is introduced into HUG1, as shown in Figures 3g and 3h, it appears that the "almost SPE" equilibrium vanishes, and that of the initial states considered lead, in the end, to the no hold-up equilibrium.

We conclude from these exercises that (a) *both strong or sunk cost sensitive and weak fairness norms can, in isolation, take root under some conditions, and so resolve the hold-up problem*, (b) *but the strong norm is less robust (and thus will be observed less often) than the weak one*, (c) *whether or not either norm is*

established, *investment rates tend to rise or fall with the costs of investment*, but (d) *the responsiveness of proposers to these costs varies with how well established these norms are*.

4 The Fairness Norms in Combination

If weak fairness is indeed the more robust of the two norms, it becomes important to determine to what extent "spilt milk" matters in environments where both norms can, in principle, co-exist. In more evocative terms, will those who do not understand Mankiw's "deep truth" be driven to (near) extinction when other, less restrictive, definitions of fairness exist? To this end, we first consider the evolution of shares in HUG3 in the absence of drift, as illustrated in Figures 5a and 5b, where the horizontal axis is still the proportion of *As* who invest, now defined as $p_{SF}^A + p_{WF}^A + p_S^A$, and the vertical is the proportion of *Bs* who are fair in *either* sense, $p_{SF}^B + p_{WF}^B$. As before, there are solution paths in which the no hold-up problem is solved but, for all those pictured here, weak fairness dominates, even if there are Nash components in which this is not the case.

[Insert Figures 5a through 5h About Here]

If a small amount of neutral or uniform drift is introduced, then the unique stable equilibrium, as reported in Table 2, is one in which almost 80% of the *As* do not invest when costs are low, and more than 90% do not when costs increase. Of those *As* who do invest, most but not all are prepared to accept even a selfish offer, but of those who would turn such an offer down, about half (5.2%, or 1.14% of all *As*) would do so because it violated the stricter norm when costs are low, a proportion that falls to less than 40% (0.95%) when costs rise.

[Insert Table 2 About Here]

An increase in the turnover rate for the *As* was sufficient in HUG1 to allow the weak fairness norm to take root for some initial conditions, but as Figures 5e and 5f reveal, this is not the case when both norms are present. In one sense, then, the presence of even small numbers of sunk cost sensitive investors undermines the position of those with less restrictive norms. It is important to note, however, that even when the hold-up problem is not resolved, almost half (!) of all *As* will invest when costs are low, and that one third will continue to do as costs increase. Furthermore, the proportion of those who invest and insist on some sort of fair offers falls a little, from 16.3% to 14.0%, as costs rise, as does the relative share of those with the sunk cost sensitive norm. On the other side of the match, the *Bs* will once more seem less opportunistic in cases where the *As* are reluctant to invest and, as illustrated below, more opportunistic when there is no reluctance: the proportion of *Bs* who would, if afforded the chance, propose selfish distributions falls as investment costs rise, from 86.5% to 72.3%.

If the turnover rates remain unequal but drift becomes prosocial, however, a second stable, no hold-up, equilibrium is established. (In the context of HUG3, we define prosocial drift as $d_{SF}^A = d_{WF}^A = 0.40$ and $d_S^A = d_D^A = 0.10$, which means that while 80% of all newcomers adopt one of the fairness norms, neither is favored over the other.) The most remarkable feature of this equilibrium is how prominent, indeed important, the sunk cost sensitive *As* become. In the case when investment costs are low, for example, almost all (97.5%, to be precise) of the *As* will invest and, of those who do, more than a third would turn down an offer that did allow them to recoup these costs and split the net surplus. It is for this reason that more than 90% of all proposed distributions meet this stricter test, an outcome that leaves the *As* better off, and the *Bs* worse off, than if the weak norm had dominated. So even if the presence of sunk cost sensitive investors is an impediment of sorts to those with weak(er) fairness norms, it nevertheless benefits them if and when the hold-up problem is resolved. As the cost of investment rises, the proportions of *As* who do not invest, who invest and would accept selfish offers and who invest and insist on weak fairness all increase at the expense of the sunk cost sensitive, but because the effects are small, the *Bs* do not become much more opportunistic: the proportion whose proposals are strictly fair falls less than five percent, from 93.7% to 88.8%. As Figures 5g and 5h remind us, however, the basin of attraction for the no hold up equilibrium becomes smaller as costs increase.

We conclude that (a) *under some conditions, the weak and sunk cost sensitive norms co-exist in some environments*, (b) *whether or not the hold-up problem is resolved, investment rates will tend to rise or fall with the costs of investment, but the opportunism or proposers varies with the level of investment*, and (c) *despite the robustness of the weak fairness norm, those who hold it sometimes benefit from the presence of the sunk cost sensitive one*.

5 Conclusion

We conclude with a caveat or two about one of the most important manifestations of hold-up outside the experimental lab, the decision not to invest, or perhaps underinvest, in firm-specific human capital (Malcomson 1997). Some readers will be tempted to frame our results as a parable about the evolution of fairness in the workplace, as an alternative to sometimes complicated contracts: workers trust that firms, or perhaps vice versa, will compensate them for the acquisition of non-portable skills and are prepared to invest in this sort of human capital because violations of the fairness norm that is the basis for such trust are punished from time to time.

This inference is premature, however. Within the framework of the model itself, for example, the existence of a no hold-up equilibrium is not assured, and even when such an equilibrium does exist, there is often another stable equilibrium in which workers do not invest. Even in the best case scenario, then, the establishment of a fairness norm, sunk cost sensitive or not, is not inevitable, and alternative remedies, not least state-sponsored ones, will some-

times be required. Furthermore, the need for such alternatives rises with the costs of investment to the extent that it becomes harder for fairness norms to "take root."

No less important, perhaps, our simple model best describes labor markets in which the worker-firm relationship is short-lived, since the random matches are formed, and then dissolved, each period. A more elaborate model, and an obvious direction for future research, would consider multi-period matches, recognizing that the long(er) time horizon also allows for the implementation of more complicated contracts (MacLeod and Malcomson 1993). It also remains to be seen how well the comparative statics properties of our model predict experimental outcomes.

6 References

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

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

Carmichael, L. and W. B. MacLeod. 2002. Caring about sunk costs: a behavioral solution to the hold-up problem with small stakes, mimeo.

Carpenter, J. P. and P. H. Matthews. 2003. No switchbacks: rethinking aspiration-based dynamics in the ultimatum game, *Middlebury College Working Paper* 02-18R.

Ellingsen, T. and J. Robles. 2002. Does evolution solve the hold-up problem? *Games and Economic Behavior*, forthcoming.

Grout, P. 1984. Investment and wages in the absence of binding contracts: a Nash bargaining approach, *Econometrica* 52: 449-460.

Güth, W., R. Schmittberger and B. Scwhwarze. 1982. An experimental analysis of ultimatum bargaining, *Journal of Economic Behavior and Organization* 3: 367-388.

Hackett, S. C. 1993. Incomplete contracting: a laboratory experimental analysis, *Economic Inquiry* 31: 274-297.

MacLeod, W. B. and J. M. Malcomson. 1993. Investments, holdup and the form of market contracts, *American Economic Review* 83: 811-837.

Malcomson, J. M. 1997. Contracts, hold-up, and labor markets. *Journal of Economic Literature* 35: 1916-1957.

Mankiw, N. G. 1998. *Principles of Microeconomics*. New York: Dryden.

Oosterbeek, H., J. Sonnemans and S. van Velzen. 1998. Bargaining with endogenous pie size and disagreement points: a holdup experiment, *Journal of Population Economics*, forthcoming.

Phillips, O. R., R. C. Battalio and C. Kogut. 1991. Sunk and opportunity costs in valuation and bidding, *Southern Economic Journal* 58: 112-128.

Schlag, Karl H. 1994. Why imitate, and if so, how? Exploring a model of social evolution, *University of Bonn Discussion Paper* B-296.

Thaler, R. H. 1980. Toward a positive theory of consumer choice, *Journal of Economic Behavior and Organization* 1: 39-60.

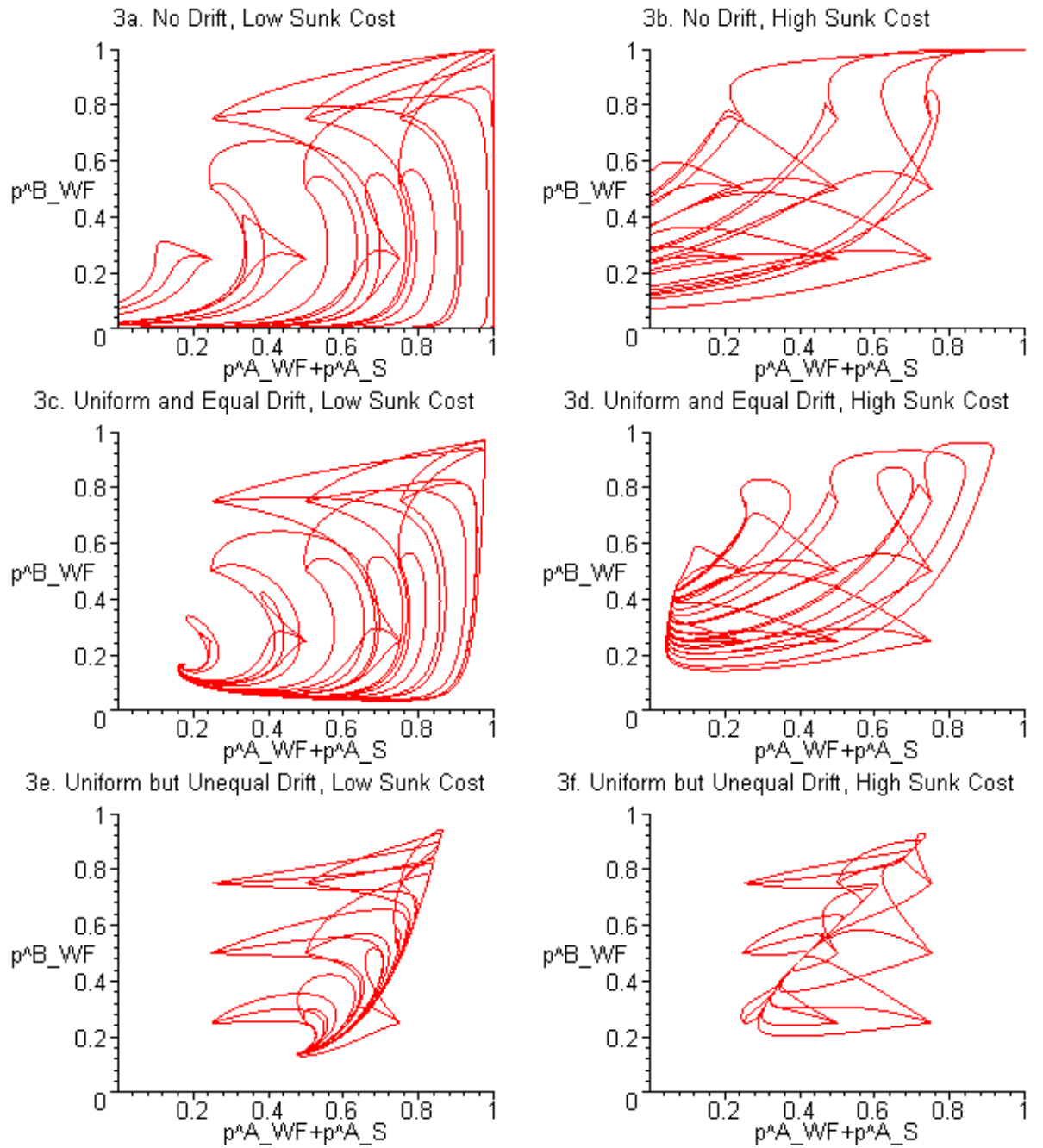
Williamson, O. 1985. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

		<i>B</i>	
		Propose Weakly (Strongly) Fair Offer	Propose Selfish Offer
<i>A</i>	Invest, Accept Only Weakly (Strongly) Fair	$2 - c, 2$ $2 - (c/2), 2 - (c/2)$	$-c, 0$
	Invest, Accept Selfish Offer or Better	$2 - c, 2$ $2 - (c/2), 2 - (c/2)$	$1 - c, 3$
	Don't Invest	$0, 0$	$0, 0$

Figure 1. The Normal Forms for HUG1 and HUG2 (Note: Where the payoffs for HUG1 and HUG2 differ, the latter are in bold.)

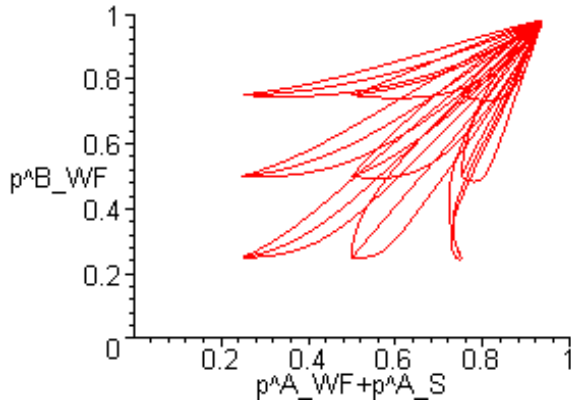
		<i>B</i>		
		Propose Strongly Fair Offer	Propose Weakly Fair Offer	Propose Selfish Offer
<i>A</i>	Invest, Accept Only Strongly Fair	$2 - \frac{c}{2}, 2 - \frac{c}{2}$	$-c, 0$	$-c, 0$
	Invest, Accept Only Weakly Fair	$2 - \frac{c}{2}, 2 - \frac{c}{2}$	$2 - c, 2$	$-c, 0$
	Invest, Accept Selfish or Better	$2 - \frac{c}{2}, 2 - \frac{c}{2}$	$2 - c, 2$	$1 - c, 3$
	Don't Invest	$0, 0$	$0, 0$	$0, 0$

Figure 2. The Normal Form for HUG3

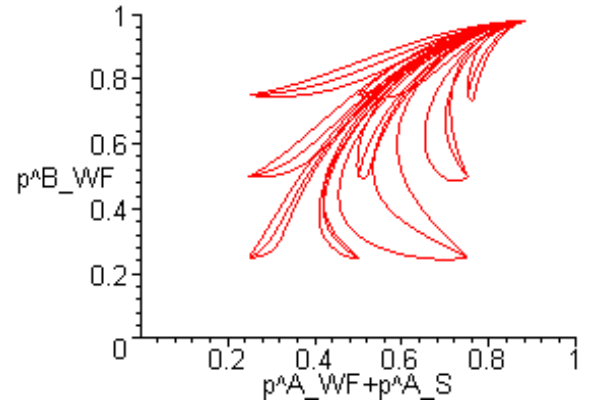


Figures 3a-3f Pseudo Phase Plots For HUG1

3g. Prosocial and Unequal Drift, Low Sunk Cost

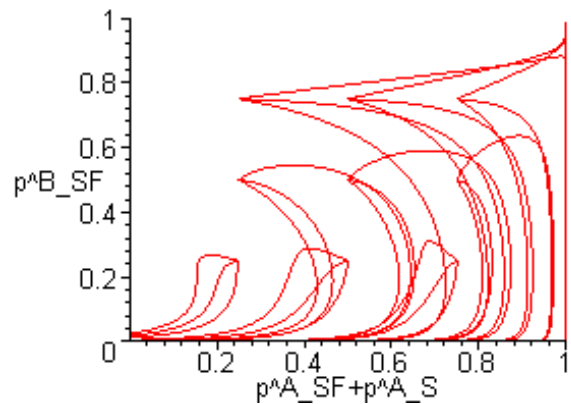


3h. Prosocial and Unequal Drift, High Sunk Cost

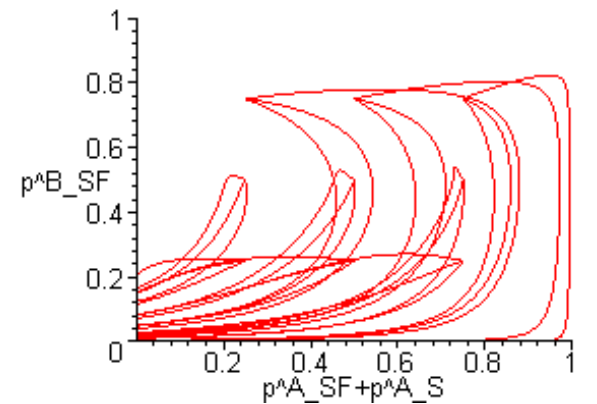


Figures 3(g) and 3(h) Pseudo Phase Plots for HUG1 (Continued)

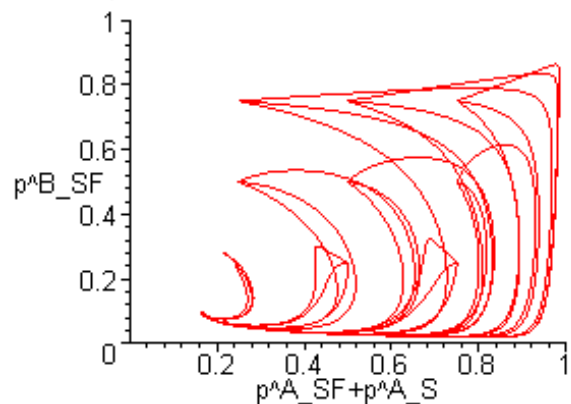
4a.. No Drift, Low Sunk Costs



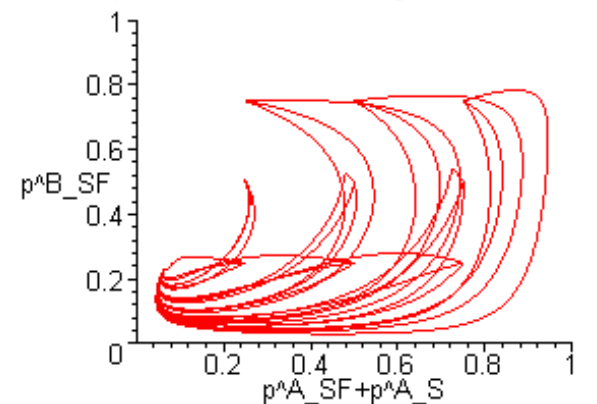
4b.. No Drift, High Sunk Costs



4c.. Equal and Neutral Drift, Low Sunk Costs

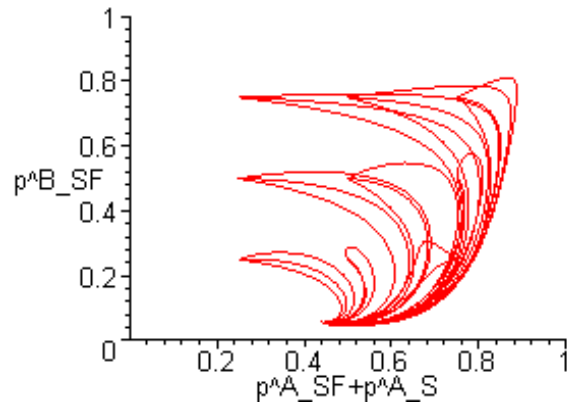


4d.. Equal and Neutral Drift, High Sunk Costs

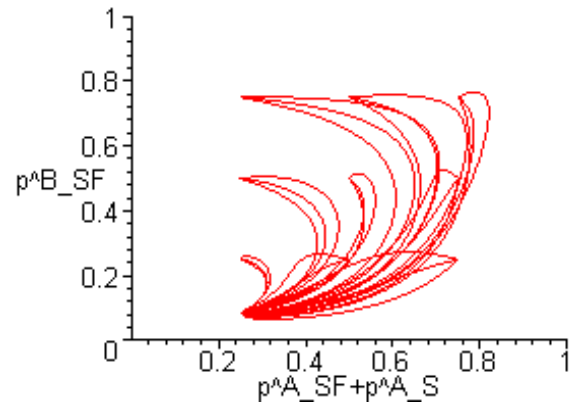


Figures 4a through 4d Pseudo Phase Plots for HUG2

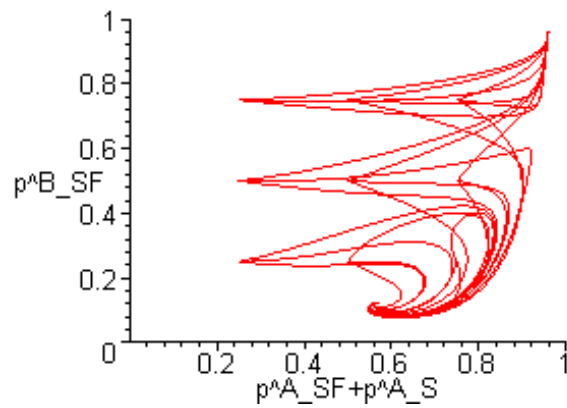
4e.. Unequal but Neutral Drift, Low Sunk Costs



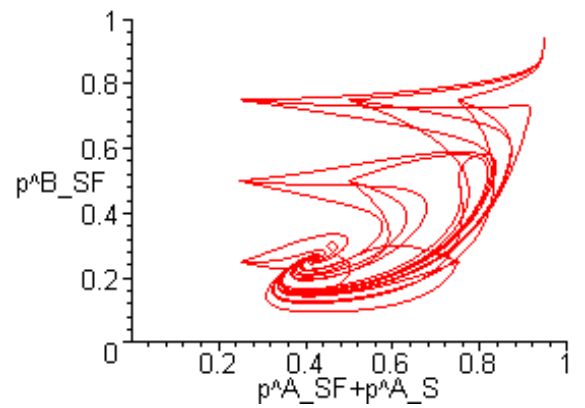
4f.. Unequal but Neutral Drift, High Sunk Costs



4g.. Unequal and Prosocial Drift, Low Sunk Costs

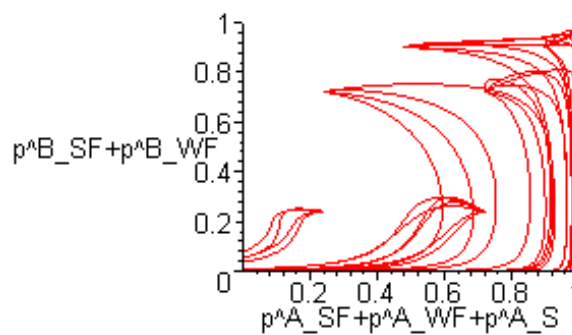


4h.. Unequal and Prosocial Drift, High Sunk Costs

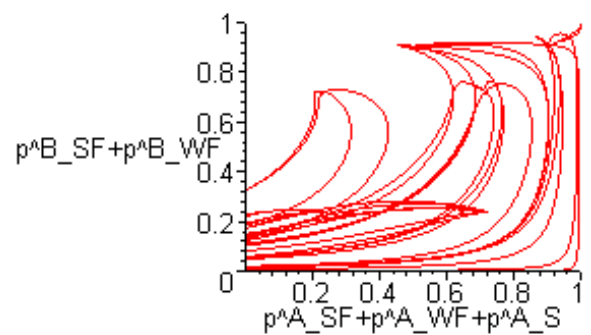


Figures 4e through 4h. Pseudo Phase Plots for HUG2 (Continued)

5a. No Drift, Low Sunk Costs

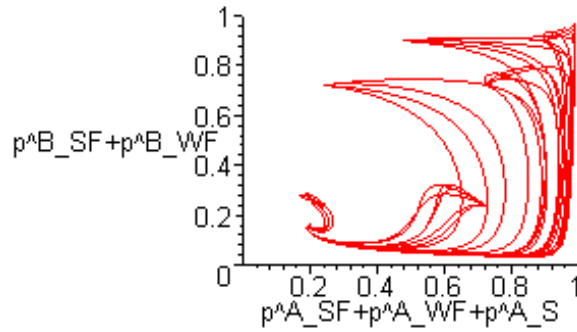


5b. No Drift, High Sunk Costs

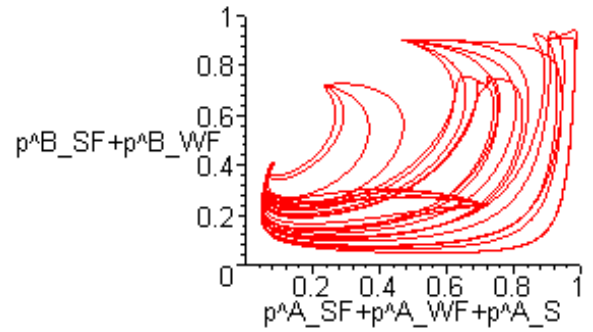


Figures 5a and 5b. Pseudo Phase Plots for HUG3

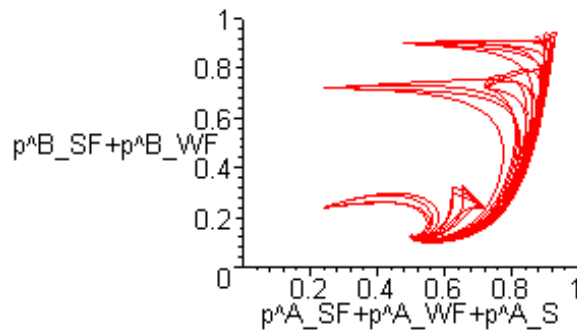
5c. Equal and Neutral Drift, Low Sunk Costs



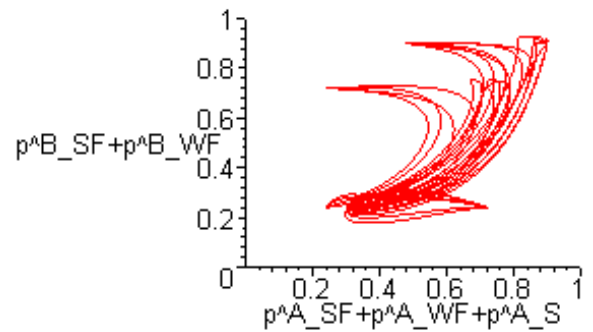
5d. Equal and Neutral Drift, High Sunk Costs



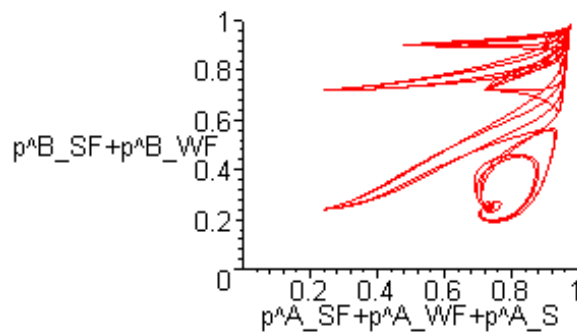
5e. Unequal but Neutral Drift, Low Sunk Costs



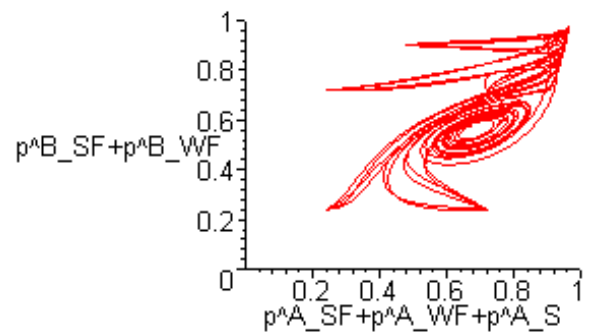
5f. Unequal and Neutral Drift, High Sunk Costs



5g. Unequal and Prosocial Drift, Low Sunk Costs



5h. Unequal and Prosocial Drift, High Sunk Costs



Figures 5c through 5h. Pseudo Phase Plots for HUG3 (Continued)

		HUG1		HUG2	
		$c = 1.25$	$c = 1.75$	$c = 1.25$	$c = 1.75$
$\theta^A = 0.01$	p_{SF}^A			1.67	1.52
$\theta^A = 0.01$	p_{WF}^A	1.79	1.88		
<i>Neutral Drift</i>	p_S^A	15.9	5.07	15.3	5.00
	p_D^A	82.3	93.1	83.0	93.4
	p_{SF}^B			9.89	23.0
	p_{WF}^B	16.5	43.6		
	p_S^B	83.5	56.4	90.1	77.0
$\theta^A = 0.075$	p_{SF}^A			10.0	8.11
$\theta^B = 0.01$	p_{WF}^A	11.0/36.8	13.0/30.9		
<i>Neutral Drift</i>	p_S^A	36.8/49.0	26.7/40.7	33.9	17.8
	p_D^A	52.2/9.39	60.3/28.4	56.1	74.1
	p_{SF}^B			5.74	9.32
	p_{WF}^B	14.2/90.6	46.7/89.4		
	p_S^B	85.8/9.39	55.3/10.6	94.3	90.7
$\theta^A = 0.075$	p_{SF}^A			24.8/75.1	24.2/73.3
$\theta^B = 0.01$	p_{WF}^A	74.2	70.1		
<i>Prosocial Drift</i>	p_S^A	19.5	18.5	32.1/20.8	17.8/21.6
	p_D^A	6.28	11.4	43.1/4.11	58.0/5.08
	p_{SF}^B			12.1/96.5	26.1/94.3
	p_{WF}^B	98.1	98.0		
	p_S^B	1.93	2.03	87.9/3.51	73.9/5.67

Table 1. Stable Population Compositions (In Percent) for HUG1 and HUG2

		$c = 1.25$		$c = 1.75$	
$\theta^A = 0.01$	p_{SF}^A	1.14		0.95	
$\theta^A = 0.01$	p_{WF}^A	1.38		1.56	
<i>Neutral Drift</i>	p_S^A	17.8		5.75	
	p_D^A	79.7		93.1	
	p_{SF}^B	6.11		15.6	
	p_{WF}^B	9.58		25.2	
	p_S^B	84.3		59.2	
	p_D^B				
$\theta^A = 0.075$	p_{SF}^A	7.62		6.07	
$\theta^B = 0.01$	p_{WF}^A	8.67		7.96	
<i>Neutral Drift</i>	p_S^A	34.2		18.3	
	p_D^A	50.49		67.7	
	p_{SF}^B	4.79		8.03	
	p_{WF}^B	8.67		19.7	
	p_S^B	86.5		72.3	
	p_D^B				
$\theta^A = 0.075$	p_{SF}^A	14.0	36.7	12.4	31.8
$\theta^B = 0.01$	p_{WF}^A	18.9	48.3	27.8	51.6
<i>Prosocial Drift</i>	p_S^A	40.3	12.5	28.2	13.5
	p_D^A	26.8	2.50	31.6	3.10
	p_{SF}^B	9.35	93.7	20.1	88.8
	p_{WF}^B	14.9	5.32	35.8	9.77
	p_S^B	75.8	0.98	44.1	1.43
	p_D^B				

Table 2. Stable Population Compositions (In Percent) for HUG3