European Journal of Economic and Social Systems
Vol. 15, N°3 (2001) 153–169

HOLDUP AND THE EVOLUTION OF BARGAINING CONVENTIONS*

Herbert Dawid[1] and W. Bentley MacLeod[1]

Abstract. As Posner (1997) has observed, when individuals in a relationship can commit to imposing costs upon each other then efficient behavior in the absence of law is possible. The question is whether efficient norms of behavior evolve endogenously in a population. We show that in a standard hold up model in which both parties make relationship specific investments the long run outcome of a stochastic adaptation process similar to Young (1993)'s 'adaptive play' does not in general correspond to the efficient equilibria. As Grossman and Hart (1986) observe, institutions, such as firms, may be needed to improve the allocation of resources.

Classification Codes. C78, L41.

1. Introduction

Consider two researchers who write a grant proposal together, where one of them writes the theoretical part of the proposal and the other one the empirical part. Let us assume that once the proposal is submitted both researchers can correctly estimate the amount of effort invested by the partner in preparing his part of the proposal (*e.g.* the investigators might use paid research assistants to prepare their part of the proposal). After the evaluation of the proposal a sum of money is assigned to the project according to its perceived quality where the allocation between the investigators is at their own discretion. Each researcher can increase the expected amount assigned to the project by investing additional effort but ex ante it is not clear which fraction of this additional revenue he will receive. In particular, under an 'equal split' rule, where the amount is equally divided

Keywords and phrases: Hold up problem, bargaining, evolution.

* *This research was supported by the National Science Foundation under grant SES-0095606.*

[1] Department of Economics, The University of Southern California, Los Angeles, CA 90089-0253, USA.

between the two main investigators, a researcher gets only half of the additional expected revenue he is generating and therefore will invest less effort than would be optimal in order to maximize the expected profit of this enterprise.

This is an instance of the well-known *hold-up* problem, which arises in scenarios where investments are sunk at the time the surplus of an interaction occurs and there are no *ex-ante* contracts to determine the allocation of this joint surplus under all contingencies. The hold-up problem has been identified as a reason for ex-ante under-investment in many individual and inter-firm relationships when the ex-post allocation of the joint surplus follows a cooperative bargaining solution like the Nash bargaining solution (*e.g.* Grout, 1984; Grossman and Hart, 1986; Klein *et al.*, 1978).

The problem is that in the absence of an agreement rational individuals ignore sunk costs when bargaining, and hence the *ex post* division of the surplus is insensitive to the level of investment. Now suppose in our example a researcher who realizes that he has invested more effort in the proposal than his partner threatens to block the project if his part of the grant is not sufficiently large to reimburse him for his additional effort. If such behavior is the norm, and this is commonly known, then it is equilibrium behavior to adhere to the norm (and accept less than the equal split if the effort was below average) and the problem of inefficient investment incentives disappears.

Beginning with Schelling (1980), it is well understood that the ability to commit to carry out threats can alter the bargaining game, and in some cases induce efficiency[3]. This type of behavior can be formally modelled with the use of the Nash demand game. After investments have been made each party makes a take it or leave it offer, and if the offers are inconsistent the game ends with no trade. In this case, any *ex post* division of the surplus is a Nash equilibrium. Carmichael & MacLeod (1997) show that under the appropriate conditions, not only does an efficient equilibrium exist, moreover when there is sufficient diversity in preferences there is a unique efficient equilibrium. This paper asks if such efficient equilibria are stochastically stable in a model of evolutionary learning.

Previous work by Tröger (2002) and Ellingsen & Robles (2002) have shown that when only one party makes a specific investment, followed by the Nash demand game then all stochastically stable equilibria are efficient. In contrast, we find that this result does not extend to the case in which both parties make complementary relationship specific investments. This negative result is useful because it helps us delineate the conditions under which legal institutions, such as contracts and firms, enhance the efficiency of exchange[4]. In situations where efficient social norms can evolve there is little need for the creation of law, and in some cases, as Ellickson (1991) has illustrated, efficient norms that are inconsistent with existing legal rules may evolve.

[3] This point is further elaborated by Frank (1988) and Hirshleifer (1987).

[4] Posner (1997) argues that it is precisely in these cases that the law plays an potentially important role. See also Mahoney and Sanchirico (2000) for a formal illustration this point in the context of 2 × 2 normal form games.

Our results suggest that this is particularly important in relationships that entail investment by both parties, a case that is central to the Hart and Moore (1990) theory of the firm. The result may also contribute to the recent debate on the role of contractual incompleteness for the theory of the firm. Maskin and Tirole (1999) have shown in a general mechanism design framework that it is possible to implement the first best investment level even though contracts are incomplete. Our analysis suggests that the norms of behavior that can implement such behavior are not necessarily stochastically stable in the case of two sided investments. It would be worthwhile to explore the stability of the mechanisms proposed by Maskin and Tirole (1999)[5].

Our analysis builds upon Young (1993b)'s model, where individuals are assumed to sample examples of past play and use this sample to guide them in the choice of play for the next period. We find that the requirement of stochastic stability, rather than improving allocation efficiency, results in a worse outcome relative to the investment levels when individuals use the Nash bargaining rule. This result is surprising given that for the Nash demand game Young (1993b) has shown that the Nash bargaining rule is stable in the absence of specific investments. Moreover, when individuals are assumed to make small mistakes in static choices, then as Binmore (1998) shows, one also obtains the equal division Nash bargaining solution[6]. Thus the results in this paper illustrate two points. First, the results suggest that bargaining conventions depend upon context, and, in particular one should be careful in assuming that the outcome of bargaining is independent of sunk costs, an assumption that is central to many theories of hold-up. Secondly, the endogenous evolution of bargaining behavior may entail the selection of investment conventions that are less efficient than those arising from the Nash bargaining solution.

2. The bargaining model

Consider a population of individuals who are randomly matched each period to engage in production and trade. It is assumed that the value of the relationship is a function of a relationship specific investment that each person makes at the beginning of the period. After investment and matching, each person observes the other's level of investment and then the available surplus is divided using the Nash demand game. The formal game played each period is as follows.

1. The $2n$ individuals simultaneously decide their investment levels $I \in \{H, L\}$, where the cost of investment is $c(I) = \begin{cases} c, & \text{if } I = H, \\ 0, & \text{if } I = L. \end{cases}$

[5] Cabrales (1999) has some interesting preliminary results showing that mechanisms based upon the Nash solution concept are adaptively stable, a concept that is different from the concept of stochastic stability used in this paper. His results do not extend to refinements of the Nash solution concept.

[6] Binmore (1998), Section 1.5.

2. Individuals are randomly matched into n bargaining pairs and observe each other's investment level. This determines the size of the surplus, $S_{I_iI_j}$, satisfying $S_{HH} \geq S_{LH} = S_{HL} \geq S_{LL} > 0$.
3. Individual i makes a demand conditional upon his investment level, I_i, and her partner's, I_j , denoted by $x_{I_iI_j} \in X_{I_iI_j}(k) = \{0, \alpha_{I_iI_j}, 2\alpha_{I_iI_j}, ..., k\alpha_{I_iI_j}\}$, where k is some large even number and $\alpha_{I_iI_j} = S_{I_iI_j}/k$.
4. The payoff to individual i is given by the rules of the Nash demand game:

$$U^i = \begin{cases} x^i_{I_iI_j} - c(I_i), & \text{if } x^i_{I_iI_j} + x^j_{I_jI_i} \leq S_{I_iI_j}, \\ -c(I_i), & \text{if not.} \end{cases}$$

Similarly for player j.

Throughout the analysis S_{HH} and S_{LL} are assumed fixed, while the degree of complementarity in investment, S_{LH}, and the cost of investment, c, are parameters that determine the nature of the investment problem. It shall always be assumed that having both individuals invest is the efficient allocation and hence:

$$S_{HH} - 2c \geq \max\{S_{LH} - c, S_{LL}\} \cdot \tag{1}$$

When S_{LH} is close to S_{HH} then investments are close substitutes, while they are strongly complementary when S_{LH} is close to S_{LL}. If $S_{LH} = S_{HL} = (S_{HH} - S_{LL})/2$ then investment is purely additive, and the marginal contribution of each person's investment is independent of the other's.

Stochastic stability is analyzed with a finite game, and hence demands are assumed to be taken from the finite set $X_{I_iI_j}(k)$, where k denotes the fineness of the grid. Finally, the Nash demand game incorporates the notion of retribution or revenge that Posner (1981) has observed to be an important ingredient for the evolution of social norms and conventions. In particular, it is important to realize that in the Nash demand game both players are assumed to have the ability to commit to 'leaving money on the table' and not to renegotiate once it becomes clear that there will be no trade. This threat, inherent in the Nash demand game, ensures the existence of efficient equilibria, as characterized in the following proposition (proofs for this and all the following results can be found in the appendix).

Proposition 1. *There exists a subgame perfect equilibrium entailing investment by both parties if and only if $S_{HH}/2 - c \geq 0$, while there is always a subgame perfect equilibrium that entails no-investment by both parties.*

Given that $S_{HH}/2 - c \geq 0$ is a necessary condition for investment by both parties to be efficient, then this result demonstrates that when commitments are possible, as modeled by the Nash demand game, then there always exist efficient equilibria. The proposition also shows that in addition to these efficient equilibria there are a huge number of other equilibria. Any bargaining strategy profile $\{(x^i_{LL}, x^i_{HL}, x^i_{LH}, x^i_{HH}), i \in \{1,2\}\}$ where $x^1_{I_1I_2} + x^2_{I_2I_1} = S_{I_1I_2}$, $\forall I_1, I_2 \in \{H, L\}$, *i.e.* where ex-post bargaining is efficient under all investment scenarios, may occur in a subgame perfect equilibrium, and in particular this range of equilibrium bargaining behavior also includes allocations which make it optimal for both players not to invest in the first period. The question we now ask is whether or not evolutionary dynamics selects the efficient equilibria from the plethora of equilibria possible in the model.

3. Evolutionary dynamics

To study the stability of the Nash equilibria in this model we use an evolutionary model, similar to Young (1993b), to characterize the set of stochastically stable bargaining conventions. Young (1993b) has shown that in the absence of investment the equal split rule is stochastically stable. Our model consists of a single population with no bargaining roles, and hence when both players invest (HH) or do not invest (LL), the only stable bargaining convention is an equal split (which here also is the unique equilibrium where behavior in the population is uniform). Therefore we may simplify the analysis by supposing that $x_{HH} = S_{HH}/2$ and $x_{LL} = S_{LL}/2$. In this case the strategy space of an individual is given by $A(k) = \{H, L\} \times X(k)^2$, where $X(k) = X_{HL}(k) = X_{LH}(k)$ and a typical strategy is denoted by the triple $a = (I, x_{HL}, x_{LH}) \in A(k)$.

Dynamics are introduced by supposing that individuals use their observations of previous play to decide how to play in the current period. Observing all past play is too costly, rather an individual samples m individuals in the previous period's population to update his memory. This information is used to construct an empirical offer distribution that is used to select the optimal strategy. Specifically the learning process proceeds as follows:

1. At the beginning of period t, each individual randomly samples the actions of m individuals from the previous period population. Let $\hat{p}_t^i \in P = \{0, 1/m, 2/m, ..., 1\}$ denote the fraction of individuals selecting $I_{t-1} = H$. Since the equal split is expected when in an HH or LL pairing, only those individuals that select either x_{HL} or x_{LH} are used to update one's memory, consisting of at most m data points at any time. It is assumed that the oldest data is dropped as new data is added. These distribution functions $\hat{F}_{HL}(\cdot)$ and $\hat{F}_{LH}(\cdot)$ are taken from the finite set:

$$\mathcal{F} = \{F : X(k) \to \{0, 1/m, 2/m, ..., 1\} \,|F(x) \text{ is increasing, } F(S_{LH}) = 1\} \cdot$$

   An individual's belief at time t is given by:

$$b_t^i = \left\{\hat{p}_t^i, \hat{F}_{HL}(\cdot), \hat{F}_{LH}(\cdot)\right\} \in P \times \mathcal{F}^2 = B.$$

2. With probability $\varepsilon > 0$ the individual selects a strategy randomly from $A(k)$, using a uniform distribution. We call this a *mutation.* The noise process is *i.i.d.* between individuals and periods. With probability $1 - \varepsilon$ the individual chooses her strategy a_t^i to maximize her utility given beliefs b_t^i. When indifferent over demands she chooses the largest demand, and similarly

when indifferent over investment, $I_t^i = H$ is chosen. Thus the agent's strategy is uniquely defined by her beliefs and can be written as $a_t^i = a\left(b_t^i\right)$.

3. Agents are randomly paired, and their payoffs are determined according to their strategies chosen at stage 2.

Given that an agent's action is completely characterized by her beliefs $b^i \in B$, then the state at time t is characterized by a distribution over beliefs, and therefore the state space is finite and given by:

$$\mathcal{S} = \left\{ s \in 0,1]^{|B|} | \sum_{b \in B} s_b = 1,\ ns_b \in \mathbb{N}_0\ \forall b \in B \right\} \cdot \tag{2}$$

The learning process described above defines a time homogeneous Markov process $\{\sigma_t\}_{t=0}^{\infty}$ on the state space $\mathcal{S}$ and the process is irreducible for positive ε. Hence, for $\varepsilon > 0$ there exists a unique limit distribution $\pi^*(\varepsilon)$ over $\mathcal{S}$, where $\pi_s^*(\varepsilon)$ denotes the probability of state s.

The long run behavior of the process for small mutation probabilities is described by the weight of the different states in the limit distribution as ε goes to zero. Clearly, the weight of states which are transient under the process governed only by the learning and decision making by the individuals goes to zero as ε shrinks. To see this, note that from any such state the process without mutations with positive probability leads to some other state where, in the absence of mutations, it is impossible to get back again. A limit set of the process is a set of states such that for $\varepsilon = 0$ the process does not leave the set if it starts in the limit set. Moreover all states within the set are visited with positive probability from every state in the set.

Definition 2. A set $\Omega \subseteq \mathcal{S}$ is called a limit set of the process if for $\varepsilon = 0$ the following statements hold:

$$\forall s \in \Omega,\ \Pr(\sigma_{t+1} \in \Omega | \sigma_t = s) = 1$$
$$\forall s, \tilde{s} \in \Omega\ \exists z > 0\ \Pr(\sigma_{t+z} = \tilde{s} | \sigma_t = s) > 0.$$

It is well known that for $\varepsilon = 0$ the process with probability one ends up trapped in one of the limit sets. However, it has been shown in various frameworks that the weight of whole limit sets goes to zero for $\varepsilon \to 0$ if the number of simultaneous mutations needed to get out of the set is smaller than the number of simultaneous mutations needed to get back into the set.[7] In the long run, the process spends almost all the time at absorbing sets where the number of mutations needed to get there relative to the number of mutations needed to get out is smallest. Such sets are called stochastically stable.

Definition 3. A state $s \in \mathcal{S}$ is called stochastically stable if $\lim_{\varepsilon \to 0} \pi_s^*(\varepsilon) > 0$. We say that a set is stochastically stable if all its elements are stochastically stable.

[7] See Kandori *et al.* (1993) and Young (1993a).

We are interested in understanding the properties of investment and bargaining behavior that evolves in the long run and will therefore focus on the stochastically stable sets of the process. In particular, we are interested in the evolution of bargaining and investment norms. In game theoretic contexts norms are generally seen as an equilibrium selection device. A norm corresponds to a strategy with the property that it is optimal to choose this strategy for any individual who believes that all other individuals will follow the norm. Here, we say that a bargaining norm or bargaining convention has evolved if there exists a pair of demands (x_{HL}, x_{LH}) with $x_{HL} + x_{LH} = S_{LH}$, such that all individuals have beliefs that x_{HL} is always demanded in a high-low pairing, x_{LH} is always demanded in a low-high pairing and all individuals actually make these demands if they are in high-low or low-high pairings. In other words, bargaining behavior is uniform in the population, efficient and correctly anticipated by all individuals. Furthermore, we say that a full-investment convention (no-investment convention) has evolved if all individuals choose high investment (low investment) and have beliefs $\hat{p} = 1$ ($\hat{p} = 0$).

The stochastically stable sets are a subset of the set of limit sets. To characterize the limit sets of the evolutionary process $\{\sigma_t\}$ it is important to understand the relationship between individual beliefs and the resulting investment decision. To make it easier to express this relationship the following notation will be used: $\mathcal{F}_{HL}^{H1} \subseteq \mathcal{F}$ denotes the set of beliefs $\hat{F}_{HL}^i$ which induce investment by agent i with beliefs $\hat{p}^i = 1$ (note that with $\hat{p}^i = 1$ the beliefs $\hat{F}_{LH}^i$ are irrelevant for the investment decision). $\mathcal{F}_{HL}^{L1} = \mathcal{F}_{HL} \setminus \mathcal{F}_{HL}^{H1}$ denotes the set of beliefs leading to no investment for $\hat{p}^i = 1$, and $\mathcal{F}_{LH}^{H0}, \mathcal{F}_{LH}^{L0}$ denote the corresponding sets for $\hat{p}^i = 0$. The size of these sets of course depends on the values of the parameters in the model and some of them might also be empty.

The first result shows that in every limit set of the process investment behavior is uniform in the population. On the other hand, beliefs about bargaining behavior (and therefore also bargaining behavior itself) might differ between individuals in the population. The only restriction is that all individual beliefs have to induce identical investment decisions.

Proposition 4. *All limit sets of the evolutionary process $\{\sigma_t\}$ have one of the following three forms:*

(a) *A single state $\{s\}$ with $s_b > 0$ only if $b = (1, \hat{F}_{HL}, \hat{F}_{LH})$ for some $\hat{F}_{HL} \in \mathcal{F}_{HL}^{H1}$.*

(b) *A single state $\{s\}$ with $s_b > 0$ only if $b = (0, \hat{F}_{HL}, \hat{F}_{LH})$ for some $\hat{F}_{LH} \in \mathcal{F}_{LH}^{L0}$.*

(c) *A pair of states $\{s^1, s^2\}$ where $s_b^1 > 0$ only if $b = (1, \hat{F}_{HL}, \hat{F}_{LH})$ for some $\hat{F}_{HL} \in \mathcal{F}_{HL}^{L1}$ and $\hat{F}_{LH} \in \mathcal{F}_{LH}^{H0}$. Furthermore, $s_b^2 = s_{\tilde{b}}^1$ for all $b = (0, \hat{F}_{HL}, \hat{F}_{LH})$ and $\tilde{b} = (1, \hat{F}_{HL}, \hat{F}_{LH})$ such that $s_{\tilde{b}}^1 > 0$, and $s_b^2 = 0$ for all other $b \in B$.*

The rather complicated looking type (c) limit set corresponds to a scenario where all agents hold some (potentially heterogeneous) beliefs which imply to invest for $\hat{p}^i = 0$ and not to invest for $\hat{p}^i = 1$. The agents then synchronously switch between investing and not investing. Whereas limit sets of type (a) and (b) always exist, cyclical limit sets do not exist if the degree of complementarity of investments is too large. By considering the most extreme beliefs it is easy to see that the set of beliefs $\mathcal{F}_{HL}^{L1}$ is empty if and only if $S_{LH} \leq \frac{1}{2}S_{HH} - c$, whereas the set of beliefs $\mathcal{F}_{LH}^{H0}$ is empty if and only if $S_{LH} < \frac{1}{2}S_{LL} + c$. The sets $\mathcal{F}_{HL}^{H1}$ and $\mathcal{F}_{LH}^{L0}$ are never empty. This in particular implies that for $S_{LH} < \frac{1}{4}(S_{HH} + S_{LL})$ no cyclical limit sets exist for any value of c. When investments are substitutes, $S_{LH} > \frac{1}{2}S_{HH}$, cyclical limit sets exist for all c.

In what follows the union of all limit sets of type (a) in proposition 4 is denoted by Ω_I (investment states), the union of all limit sets of type (b) by Ω_N (no-investment states) and all limit sets of type (c) by Ω_C (cyclical states).

In none of these limit sets any high-low pairings can occur without mutations, and this implies that the beliefs about bargaining behavior drift around driven by mutations. A single mutation is sufficient to lead the process into a different limit set where individuals have different beliefs. Assume the process is in some limit set and a single mutation occurs which leads to a high-low pairing. Then a pair of demands is generated, where one of the demands (that of the mutant) is completely random. With positive probability these demands are observed by some individuals who then change their beliefs $\hat{F}_{HL}$ and $\hat{F}_{LH}$. Such a change of beliefs moves the process into a different limit set and therefore it can jump from one limit set to the next in steps which need only a single mutation. Generally, a union of limit sets is called mutation connected if every state can be reached from every other state in the set in a sequence of one-mutation transitions.

Definition 5. A union of limit sets $\Omega \subseteq \mathcal{S}$ is mutation connected if for all $s, \tilde{s} \in \Omega$ there exists a sequence of states $(s_1 = s, s_2, \ldots, s_n = \tilde{s})$, $s_i \in \Omega$ such that every transition from s_i to s_{i+1}, $i = 1, \ldots, n-1$ needs not more than one mutation.

The following Lemma shows that each of the unions of the limit sets Ω_I, Ω_N and Ω_C is mutation connected and that, whenever there exist cyclical limit sets, the states of full investment and no investment are mutation connected with each other and can be reached with positive probability from every state in Ω_C.

Lemma 6.

(a) *The sets Ω_I, Ω_N and Ω_C are mutation connected.*

(b) *Whenever $\Omega_C \neq \emptyset$, i.e. whenever $S_{LH} \geq \frac{1}{2}S_{HH}$ or $\frac{1}{4}(S_{HH} + S_{LL}) \leq S_{LH} < \frac{1}{2}S_{HH}$, $\frac{1}{2}S_{HH} - S_{LH} < c \leq S_{LH} - \frac{1}{2}S_{LL}$ the union $\Omega_I \cup \Omega_N$ is mutation connected.*

(c) *Whenever $\Omega_C \neq \emptyset$, any state in $\Omega_I \cup \Omega_N$ can be reached from any state in Ω_C by one mutation steps.*

The lemma has important implications for stochastic stability, since it has been shown by Nöldecke & Samuelson (1993) that if a limit set is stochastically stable then all the limit sets which are mutation connected to this limit set are stochastically stable as well. In the following proposition the stochastically stable limit sets are characterized as a function of S_{LH} and c. The standard way to prove which of the limit sets are indeed stochastically stable has been to invoke the Markov chain tree theorem and then to examine the number of mutations needed along a large number of different directed graphs in the state space. See Young (1998) for a more thorough discussion of this technique. Sufficient conditions for stochastic stability which are much easier to verify have been recently developed in Ellison (2000), and these criteria are used in the proof here.

Proposition 7. *For sufficiently large n, m, k the long run properties of the evolutionary process $\{\sigma_t\}$ are as follows:*

(I) $S_{LH} < \frac{1}{4}(S_{LL} + S_{HH})$

There exists a threshold $c^* = \frac{1}{4}(S_{HH} - S_{LL})$ such that only Ω_I is stochastically stable for $c < c^*$ and only Ω_N is stochastically stable for $c > c^*$.

(II) $\frac{1}{4}(S_{LL} + S_{HH}) < S_{LH} < \frac{1}{2}S_{HH}$

There exists values

$$c^1 = \frac{1}{2}S_{HH} - S_{LH}, \qquad c^2 = S_{LH} - \frac{1}{2}S_{LL}$$

such that only Ω_I is stochastically stable for $c < c^1$, only Ω_N is stochastically stable for $c > c^2$ and $\Omega_I \cup \Omega_N$ is stochastically stable for $c \in [c_1, c_2]$.

(III) $\frac{1}{2}S_{HH} < S_{LH} \leq S_{HH}$

The union $\Omega_I \cup \Omega_N$ is stochastically stable for all $c \in \left[0, \frac{1}{2}(S_{HH} - S_{LL})\right]$.

These results give a rather complete characterization of the long run evolution of investment conventions. This is illustrated in Figure 1 where we show the range of cost values c leading to stable investment, stable non-investment conventions and fluctuations between the two for different values of S_{LH}.

Let us now compare the results we get under our assumption that investment and bargaining behavior evolves endogenously with those one would get in this model under the standard assumption that the joint surplus is allocated according to the Nash bargaining solution (which here simply corresponds to the equal split). The shaded area in Figure 1 consists of all pairs (c, S_{LH}) where full investment of both partners would be efficient, but under the standard assumption of an equal split allocation of the joint surplus there exists no equilibrium where both partners invest. So, the shaded region corresponds to all scenarios where the *hold-up* problem occurs according to the literature initiated by Grossman & Hart (1986) and Grout (1984). Straightforward calculations show that the shaded area is given by $S_{HH}/2 - c < S_{LH}/2$. Below the shaded area high investment is a Nash equilibrium, under Nash bargaining, while above the shaded area it is efficient for only one party to make an investment. Our main result is that the set of parameter values for which choosing the efficient investment level is stochastically stable is strictly smaller than the region where efficient investment occurs under the Nash bargaining rule. Thus, in contrast to Tröger (2002) and Ellingsen & Robles (2002), we find that in the case where both parties might invest, the requirement that norms of bargaining be stochastically stable exacerbate the hold-up problem. In the region to the right, corresponding to high S_{LH} (investments are substitutes), both high and low investment equilibria are stochastically stable. This means

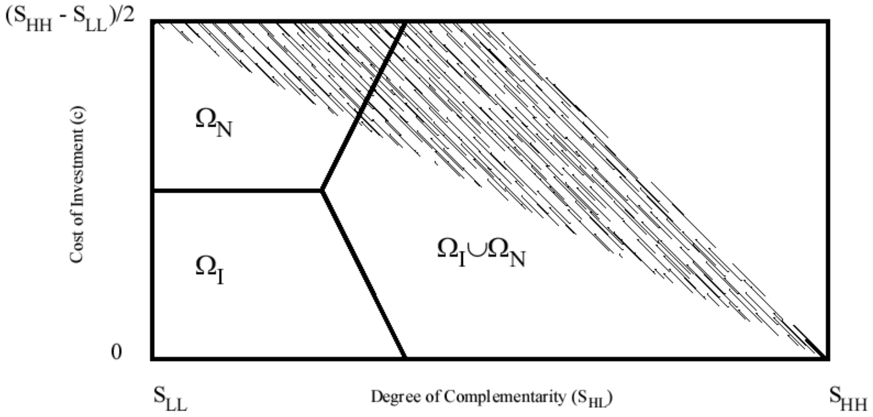

FIGURE 1. The set of parameters where full investment (Ω_I), no investment (Ω_N) and fluctuations between the two ($\Omega_I \cup \Omega_N$) are stochastically stable compared with the region where no investment (shaded area) and investment (below the shaded area) is the equilibrium outcome under the equal split rule.

that for small mutations (ε small), the system spends long periods of time at the high investment equilibrium and then long periods of time at the low investment equilibrium.

It should be noted that for none of the parameter constellations a bargaining convention is developed in the long run. Considering the different stochastically stable sets notice that even if investment is in the long run uniform, beliefs and bargaining behavior are in general heterogeneous. The reason is that once the investment strategies in the population are uniform any bargaining norm which might exist at that point will be slowly destroyed. Under uniform investment strategies the only way a pairing between a high and a low type might occur is that at least one of the two has mutated. A mutant may not follow the bargaining norm and hence at least half of the demands in high low pairings are completely random and do not follow any bargaining norm. Since all individuals use these demands to update their beliefs any uniform consistent point beliefs which might have existed in the population will be destroyed. Hence in this model stable norms of bargaining cannot evolve[8].

Given that, it becomes apparent that an investment convention can only be stochastically stable if the drift of beliefs about bargaining behavior can never change the optimal investment decision. We only get a stable full investment

[8] This analysis suggests to consider a model where the effect of investment is uncertain and therefore high-low pairings might occur with positive probability even under uniform investment. In such a model there is selection pressure on the bargaining behavior also after investment behavior has become uniform. Dawid & Macleod (2002) show that in such a model bargaining conventions are developed in the long run.

convention if high investment is optimal for all possible beliefs $\hat{F}_{HL} \in \mathcal{F}$ as long as $\hat{p} = 1$ [9]. This however is much more restrictive than the condition that high investment is optimal under the beliefs that bargaining follows the equal split rule. Hence the inability to develop bargaining norms is clearly detrimental for the evolution of full investment conventions.

4. Stability of norms

As Posner (1997) concludes from the work of Trivers (1971) and Hirshleifer (1987), when individuals can commit to walking away from profitable exchange, then in principle agreements between individuals can be made self-enforcing (see the discussion following Proposition 1 above). The question then is, given the potential for mutual harm, will efficient conventions of behavior evolve. In this paper we have explored this question in the context of hold-up when individuals make complementary investments. The ability to commit to harmful actions follows the literature and uses the Nash demand game to model bargaining behavior. In this game efficient equilibria always exist, as suggested informally by Posner (1997). We find that for a large range of parameter values, not only are these efficient equilibria not stochastically stable, but that the stable equilibria entail worse outcomes than if commitment were not possible, as is assumed in the hold-up models of Grossman & Hart (1986) and Grout (1984).

Thus we find that stable efficient norms do not always arise in a population where bargaining and investment behavior is endogenously developed even when efficient equilibria exist. Hence, in general evolution does not 'solve' the hold up problem as suggested by the results for the case of one-sided investment.

More generally our findings show that the difficulty of establishing norms follows from the fact that acting according to the norm is never required as long as all individuals believe that the rest of the population will follow the norm. Whereas this general point has been made in the literature on the evolutionary stability of backward induction (see *e.g.* Nöldecke & Samuelson, 1993 or Binmore *et al.*, 1998) our model highlights the negative efficiency implications of this effect.

Finally, notice that the inability to create the right bargaining norm is not the only source of inefficiency in this evolutionary framework. Assume that due to renegotiations or externally imposed bargaining norms the joint surplus is indeed always allocated by the equal split rule. Given this allocation there exists a high investment equilibrium for the entire region below the shaded set in Figure 1. In Dawid & MacLeod (2002) it is shown that while the set of parameters with stochastically stable high investment equilibria is strictly larger than in the model considered in this paper, it is still smaller than the entire set below the shaded area. Hence allowing for renegotiation increases the set of conditions under which bargaining conventions are stable and efficient, but even if there is a bargaining convention in place which allows for a high investment equilibrium the population

[9] Note that this is equivalent to $\mathcal{F}_{HL}^{L1} = \emptyset$.

might fail to coordinate on this efficient equilibrium (see *e.g.* Young, 1998 for a discussion of stochastic stability in coordination games).

Appendix

Proof of Proposition 1. A strategy in this game is described by an investment decision and demands. Observe that since individuals cannot condition their strategies on their roles, efficiency entails unique demands in the case of both investing ($x_{HH} = S_{HH}/2$) and neither investing ($x_{LL} = S_{LL}/2$). Efficiency also entails $x_{HL} + x_{LH} = S_{LH}$ and hence individual i will invest if and only if

$$S_{HH}/2 - c \geq S_{LH} - x_{HL} = x_{LH}. \tag{3}$$

Given that any division of the gains from trade forms a Nash equilibrium, the maximal punishment that individual j can impose upon individual i for not investing is to set $x^P_{HL} = S_{LH}$, from which one concludes that high investment by both parties is a subgame perfect Nash equilibrium if and only if $S_{HH}/2 - c \geq S_{LH} - x^P_{LH} = 0$.

Similarly, in the case of both parties not investing, individual i can deter individual j from investing by setting $x^N_{LH} = S_{LH}$ (and hence $x^N_{HL} = 0$). □

Proof of Proposition 4. First, it is easy to see that all three types of sets are absorbing for the Markov process with $\varepsilon = 0$. In all three cases the beliefs of the agents induce identical investment decisions which implies that a match between a high and a low type can never be observed and the beliefs are never changed. In cases (a) and (b) the induced investment decision always coincides with the current type, in case (c) it is always opposite and agents cycle between investment and no investment. Second, we have to show that for any other state there exists a positive probability to reach one of the limit sets. Assume that s_t is an arbitrary state where both high and low types are present. We first show that there is a positive transition probability to a state where all agents are identical in type and beliefs. Note that as long as the process has not reached any of the limit sets at least every third period there have to be both high and low types in the population (in case all agents invest in two consecutive periods the beliefs must be such that all agents invest for $\hat{p} = 1$ and the process is in a limit set of type (a), in case all agents do not invest for two consecutive periods the process must be in a limit set of type (b), and if all agents synchronously switch twice between high and low investment the process must be in a limit set of type (c)). Hence, there is a positive probability that at least m high-low pairings occur within the following $3m$ periods and that all individuals observe the same m high-low pairings. In such a case beliefs $\hat{F}_{LH}$ and $\hat{F}_{HL}$ are homogeneous in the population. With positive probability all agents then observe an identical fraction of high types which yields a homogeneous population.

Now, we have to show that there is a positive probability to reach one of the sets listed in (a), (b) and (c) from an arbitrary homogeneous state. Assume without loss of generality that all agents are of type H. We denote the identical expectations of all agents by $\tilde{F}_{HL}$ and $\tilde{F}_{LH}$. If $\tilde{F}_{HL} \in \mathcal{F}_{HL}^{H1}$ we have reached a state of type (a). If $\tilde{F}_{HL} \in \mathcal{F}_{HL}^{L1}$ all agents choose not to invest and in the following period all agents are of type L. If $\tilde{F}_{LH} \in \mathcal{F}_{LH}^{L0}$ we have reached a state of type (b) and if $\tilde{F}_{LH} \in \mathcal{F}_{HL}^{H0}$ all agents will invest again in this period and we have reached a cycle of type (c). In any case one of the limit sets is reached. □

Proof of Lemma 6. We prove (a) only for Ω_I. It can be shown for the other sets using similar arguments. Take an arbitrary state $s \in \Omega_I$ and another arbitrary state $s' \in \Omega_I$. The following transition from s to s' needs only one mutation in each step from one limit set in Ω_I to the next. Let $X'_{i,HL}$ and $X'_{i,LH}$ denote the set of observations of values x_{HL} and x_{LH} agent i has to make such that the resulting empirical distribution functions of all individuals yield the distribution given in s'. We show that in one-mutation steps first all the observations in $X'_{i,HL}, i = 1, \ldots n$ and then all the observations in $X'_{i,LH}, i = 1, \ldots n$ can be generated. Whenever a bid $\tilde{x}_{HL} \in X'_{i,HL}$ or $\tilde{x}_{LH} \in X'_{i,LH}$ is generated in the population there is a positive probability that this bid is only observed by individual i. Furthermore, there is a positive probability that all other observations made by any individual in this period involve only bids in high-high pairings and hence do not alter $\hat{F}^i_{HL}$ or $\hat{F}^i_{LH}$. Hence, such a transition leads from s to s' in one-mutation steps. In order to guarantee that such a transition never leads out of Ω_I the memory of all agents might first be filled with observations $x_{HL} = S_{LH}$ before the observations required in s' are added.

Consider an arbitrary state $\tilde{s} \in \Omega_I$ and let $\tilde{x}_{HL}$ be an arbitrary element of $X'_{i,HL}$ for some i. Assume that one agent mutates and chooses not to invest and to bid $x_{LH} = S_{LH} - \tilde{x}_{HL}$. Assume further that only one agent – we call this agent a – observes this mutated bid and all other observations made this period are of bids in high-high pairings. This single mutation event leads from the limit set $\{\tilde{s}\}$ to a limit set $\{\tilde{\tilde{s}}\}$ where only the beliefs $\hat{F}^a_{LH}$ of agent a have been changed (note that changes in $\hat{F}_{LH}$ can never lead out of Ω_I). In the same way $m-1$ additional one-mutation steps can lead to a limit set where all beliefs but $\hat{F}^a_{LH}$ are identical to those in $\tilde{s}$ and $\hat{F}^a_{LH} = \mathcal{P}(S_{LH} - \tilde{x}_{HL})$, where $\mathcal{P}(z)(\cdot)$ denotes the distribution function of point expectations z, *i.e.* $\mathcal{P}(z)(x) = 0$ for $x < z$ and $P(z)(x) = 1$ for $x \geq z$. If there is one mutation in this limit set where one other agent mutates to a low type and is matched with agent a, a bid $x_{HL} = \tilde{x}_{HL}$ can be observed. Note that the generation of this bid $\tilde{x}_{HL}$ does not need any change of the beliefs $\hat{F}_{HL}$ of any of the agents. An arbitrary bid $\tilde{x}_{LH}$ can be generated directly by a single mutation of some agent to invest low and bid $x_{LH} = \tilde{x}_{LH}$. Using our arguments above this shows that s' can be reached from s via a path in Ω_I which needs only one-mutation steps.

To establish (b) we show that a state in Ω_N can be reached from any state in Ω_I using one-mutation transitions. Denote by $\bar{x}_{HL}$ the smallest $x_{HL} \in X$ such that agents with such point beliefs would still invest in a population with full investment, *i.e.* such that

$$S_{LH} - x_{HL} \leq \frac{S_{HH}}{2} - c.$$

Since $\Omega_C \neq \emptyset$ we have $\bar{x}_{HL} > 0$. This implies that there are beliefs $\bar{F}_{HL} \in \mathcal{F}^{H1}_{HL}$ such that adding a single observation of $x_{HL} = 0$ leads to beliefs $\bar{\bar{F}}_{HL} \in \mathcal{F}^{L1}_{HL}$. Since Ω_I is mutation connected we can get with one-mutation steps to a state in Ω_I where all agents have beliefs $\hat{F}_{HL} = \bar{F}_{HL},\ \hat{F}_{LH} = \mathcal{P}(S_{LH})$. Clearly, $\mathcal{P}(S_{LH}) \in \mathcal{F}^{L0}_{LH}$. If one of the agents now mutates to a low type, a bid $x_{HL} = 0$ can be observed. There is a positive probability that this bid is observed by all agents changing the beliefs to $\hat{F}_{HL} = \bar{\bar{F}}_{HL} \in \mathcal{F}^{L1}_{HL}$ for all individuals. The resulting state is in Ω_N and we have shown that this transition needs only one-mutation steps. Similar arguments show that also the transitions from Ω_N to Ω_I can occur via paths through limit sets which need only one-mutation per step.

For (c) consider an arbitrary state in Ω_C. Denote by $\underline{F}_{HL}$ beliefs in $\mathcal{F}^{L1}_{HL}$ such that adding a single observation of $x_{HL} = S_{LH}$ yields beliefs in $\mathcal{F}^{H1}_{HL}$. Since Ω_C is mutation connected there is a transition in single mutation steps to a state in Ω_C where all individuals have beliefs $\hat{F}_{HL} = \underline{F}_{HL}$. Assume that in a period of the cycle where all other individuals invest low one individual mutates, invests high and demands $x_{HL} = S_{LH}$. There is a positive probability that this is observed by all agents yielding beliefs $\hat{F}_{HL} \in \mathcal{F}^{H1}_{HL},\ \hat{F}_{LH} \in \mathcal{F}^{H0}_{LH}$. Hence all individuals invest high in the following period and keep investing in the following periods. A state in Ω_I has been reached. In the same way a state in Ω_N can be reached from an arbitrary state in Ω_C. □

Proof of Proposition 7. We use the radius modified coradius criterion introduced in Ellison (2000). For a union of limit sets Ω the radius $R(\Omega)$ is defined as the minimum number of mutations needed to get to a state outside the basin of attraction of Ω with positive probability. The modified coradius $CR^*(\Omega)$ is defined as follows: consider an arbitrary state $x \notin \Omega$ and a path $(z_1, z_2, \ldots, z_T)$ from x to Ω where $L_1, L_2, \ldots, L_r \subset \Omega$ is the sequence of limit sets the path goes through (this implies $L_r \subseteq \Omega$). The modified cost of this path is defined by

$$c^*(z_1, \ldots, z_T) = c(z_1, \ldots, z_T) - \sum_{i=2}^{r-1} R(L_i),$$

where $c(z_1, \ldots, z_T)$ gives the number of mutations needed on the path $(x_1, \ldots, z_T)$. Denoting by $c^*(x, \Omega)$ the minimal modified costs for all paths from x to Ω the modified coradius is given by

$$CR^*(\Omega) = \max_{x \notin \Omega} c^*(x, \Omega).$$

Ellison (2000) proves that every union of limit sets Ω with $R(\Omega) < CR^*(\Omega)$ contains all stochastically stable states.

First we show (i). In this case both $\mathcal{F}_{HL}^{L1}$ and $\mathcal{F}_{LH}^{H0}$ are empty. We will calculate the radius Ω_I and Ω_N which is the minimum number of mutations needed to get from one of these sets into the other. Let us start with the radius of Ω_I. We have to find the minimal number of mutations needed at some state in Ω_I such that an agent who enters the population decides with positive probability not to invest. It is easy to see that in such a case there is also a positive probability that all agents change to no investment in the next period and the process wanders into Ω_N. Since $\mathcal{F}_{HL}^{L1}$ is empty, non-investment can only be optimal if $\hat{p} < 1$. For a given $\hat{p}$ the incentives to change to non-investment are largest if the beliefs of all agents are such that in a match of high and low type the high type gets nothing and the low type the full amount S_{LH}. Assuming these extreme beliefs about x_{LH} and x_{HL}, an agent decides not to invest if

$$\hat{p}\frac{S_{HH}}{2} - c < (1-\hat{p})\frac{S_{LL}}{2} + \hat{p}\,(S_{LH}),$$

which gives

$$1 - \hat{p} > y_I^* := \frac{\dfrac{S_{HH}}{2} - S_{LH} - c}{\dfrac{1}{2}(S_{HH} + S_{LL}) - S_{LH}}\,.$$

Let $m_I^* = \lceil m y_I^* \rceil$ denote the minimal number of low types needed in the population such that an agent can have beliefs $\hat{p} < p_I^* := 1 - y_I^*$ after taking the sample. Hence m_I^* is the minimal number of mutations needed to get out of Ω_I if all mutations are such that agents choose being a low instead of a high type. This establishes $R(\Omega_I) = m_I^*$. For further reference we denote the initial state of this transition by $z^{I1} \in \Omega_I$ and the resulting state by $z^{N1} \in \Omega_N$.

Considering the transition from Ω_N to Ω_I, analogous arguments show that the minimal $\hat{p}$ necessary to induce investment is given by

$$\hat{p} > y_N^* := \frac{\dfrac{S_{LL}}{2} - S_{LH} + c}{\dfrac{1}{2}(S_{HH} + S_{LL}) - S_{LH}}\,.$$

and that a transition from Ω_N to Ω_I needs at least m_N^* mutations, where $m_N^* = \lceil m y_N^* \rceil$. We denote the initial and resulting state of this transition by $z^{N2} \in \Omega_N$ and $z^{I2} \in \Omega_I$.

Now consider an arbitrary $z^I \in \Omega_I$ and an arbitrary $z^N \in \Omega_N$. Since all and the limit sets in Ω_N and Ω_I are mutation connected, there has to be a path $\{z_1, \ldots z_M\}$ such that $z_1 = z^I, z_k = z^{I1}, z_{k+1} = z^{N1}, z_M = z^N$ for some k where $c(z_i, z_{i+1}) = 1\ \forall i = 1, \ldots, k-1, k+1, \ldots M-1$. Clearly we have $R(z_i) = 1\ \forall i$ and therefore the modified costs of this path are

$$c^*(z_1, \dots z_M) = M - 1 + c(z^{I1}, z^{N1}) - \sum_{i=2}^{M-1} 1 = m_I^*.$$

Since this number is independent of z_I and z_N and all limit sets which do not belong to Ω_I belong to Ω_N we get

$$CR^*(\Omega_I) = \max_{z \notin \Omega_I} c^*(z, \Omega_I) = m_I^*.$$

Analogous arguments show that $CR^*(\Omega_N) = m_N^*$. Therefore all states in Ω_I but no state in Ω_N is stochastically stable if $m_I^* > m_N^*$. For $m_I^* < m_N^*$ all states in Ω_N but no state in Ω_I is stochastically stable. For large m the inequality $m_I^* > m_N^*$ is equivalent with $y_I^* > y_N^*$ which is equivalent with

$$c < \frac{1}{4}(S_{HH} - S_{LL})$$

and we have shown (i). The other parts follow directly by taking into account Lemma 6 and the conditions under which Ω_C is non-empty. □

References

Binmore K. (1998) *Game Theory and the Social Contract II: Just Playing.* Cambridge, MA, USA: MIT Press.

Binmore K., Piccione M. and Samuelson L. (1998) Evolutionary stability in alternating-offers bargaining games. *J. Econ. Theory* **80**, 257–291.

Cabrales A. (1999) Adaptive dynamics and the implementation problem with complete information. *J. Econ. Theory* **86**, 159–184.

Carmichael L.H. and MacLeod W.B. (1997) Territorial bargaining. Technical Report 97-14, The Law School, University of Southern California.

Dawid H. and MacLeod W.B. (2002) Evolutionary Bargaining with cooperative investments. Mimeo, University of Southern California.

Ellickson R.C. (1991) *Order Without Law: How Neighbors Settle Disputes.* Cambridge, MA: Harvard University Press.

Ellingsen T. and Robles J. (2002) Does evolution solve the hold-up problem? *Games and Economic Behavior* **39**, 28–53.

Ellison G. (2000) Basins of attraction, long run stability and the speed of step-by-step evolution. *Rev. Econ. Studies* **67**, 17–45.

Frank R.H. (1988) *Passions within Reason.* New York, NY, U.S.A.: W.W. Norton & Company.

Grossman S.J. and Hart O.D. (1986) The costs and benefits of ownership: A theory of vertical and lateral integration. *J. Political Econ.* **94**, 691–719.

Grout P. (1984, March) Investment and wages in the absence of binding contracts: A nash bargaining approach. *Econometrica* **52**, 449–460.

Hart O.D. and Moore J.H. (1990) Property rights and the nature of the firm. *J. Political Econ.* **98**, 1119–58.

Hirshleifer J. (1987) The emotions as guarantors of threats and promises. In J. Dupre, ed., *The Latest on the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press.

Kandori M., Mailath G. and Rob R. (1993) Learning, mutation and long run equilibria. *Econometrica* **61**, 27–56.

Klein B., Crawford R. and Alchian A. (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J. Law and Econ.* **21**, 297–326.

Mahoney P. and C. Sanchirico (2000) Competing norms and social evolution. University of Pennsylvania Law Review.

Maskin E. and Tirole J. (1999) Unforeseen contingencies and incomplete contracts. *Rev. Econ. Studies* **66**, 83–114.

Nöldecke G. and Samuelson L. (1993) An evolutionary analysis of backward and forward induction. *Games and Economic Behavior* **5**, 425–454.

Posner R.A. (1981) *The Economics of Justice*. Cambridge, MA: Harvard University Press.

Posner R.A. (1997, May) Social norms and the law: An economic approach. *Am. Econ. Rev.* **87**, 365–369.

Schelling T.C. (1980) *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Trivers R.L. (1971) The evolution of reciprocal altruism. *Quart. Rev. Biol.* **46**, 35–49.

Tröger T. (2002) Why sunk costs matter for bargaining outcomes: An evolutionary approach. mimeo, University College London.

Young H.P. (1993a) The evolution of conventions. *Econometrica* **61**, 57–84.

Young H.P. (1993b) An evolutionary model of bargaining. *J. Econ. Theory* **59**, 145–168.

Young P. (1998) *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.