Does Evolution Solve the Hold–up Problem?

Tore Ellingsen* Jack Robles**

First version: April 1, 1999
This version: February 10, 2000

Stockholm School of Economics
Working Paper Series in Economics and Finance, No 358

Abstract

The paper examines the theoretical foundations of the hold–up problem. At a first stage, one agent decides on the level of a relationship–specific investment. There is no contract, so at a second stage the agent must bargain with a trading partner over the surplus that the investment has generated. We show that the conventional underinvestment result hinges crucially both on the assumed bargaining game and on the choice of equilibrium concept. In particular, we prove the following two results. (i) If bargaining proceeds according to the Nash demand game, any investment level is subgame perfect, but only efficient outcomes are stochastically stable. (ii) If bargaining proceeds according to the ultimatum game (with the trading partner as proposer), only the minimal investment level is subgame perfect, but any investment level is stochastically stable.

JEL classification: L14, C78.
Keywords: Specific investments, opportunism, evolution, fairness.

---

*Department of Economics, Stockholm School of Economics, Box 6501, S—113 83 Stockholm, Sweden. Email: gte@hhs.se.

**Department of Economics, University of Colorado. Email: roblesj@spot.colorado.edu.

This work was initiated while Robles was visiting the Stockholm School of Economics, for whose hospitality he is most grateful. Thanks to Paul Milgrom, Joel Sobel, Joel Watson and H. Peyton Young for helpful discussions. Financial support from Swedish Council for Research in the Humanities and Social Sciences (Ellingsen) is gratefully acknowledged.

1 Introduction

Many investments are relationship–specific. Leading examples include workers who acquire a skill that is only valued by a particular employer and sellers who develop a product which is only desired by a particular buyer. A vast recent literature builds on the notion that the specificity of the investment makes the investor vulnerable, and that this leads to underinvestment. An efficient level of relationship–specific investments requires that the investor's return is protected; contractually, through reputation, or through some other institutional arrangement. While we appreciate the importance of this literature, the purpose of the present paper is to examine critically the basic premise that unregulated bargaining offers the investor insufficient protection.

Suppose it is not possible to write any contract, neither on the investment level itself nor on the terms of future transactions. Instead, the parties bargain over the terms of trade after the investment cost has been sunk. The key idea, which was articulated quite clearly already by Klein, Crawford and Alchian (1978), is that sunk cost costs should not affect bargaining outcomes. Ex post, agents bargain over the full benefit from trade, the "quasi–rent", rather than over the benefit net of investment costs. If so, as shown by Grout (1984), Grossman and Hart (1986), and Tirole (1986), there is necessarily underinvestment, because the investing party will not be able to appropriate the full return to the (marginal) investment. This is the *hold–up problem,* or at least the most popular version of it.[1]

While the irrelevance of sunk costs is an appealing principle for single person decisions, where there is usually a unique ex post efficient outcome, it is a problematic principle for bargaining, where there is usually a large set of ex post efficient outcomes. Indeed, we shall argue that the investment cost may be highly relevant for ex post bargaining, and that plausible theories of bargaining may even admit efficient investment. The existence of a hold–up problem depends crucially both on the specific bargaining procedure and on the choice of solution concept.

To put the discussion in perspective, recall that both Grout (1984) and Grossman and Hart (1986) assume that the bargaining outcome is determined by the Nash bargaining solution, and that parties bargain over the gross surplus. Thus, sunk costs are simply not allowed to matter in their framework. There is a hold–up problem by construction. Tirole (1986) on the other hand

[1]As observed by Holmström and Roberts (1999), the hold–up terminology has come to cover a range of conceptually quite distinct ideas which have in common the notion that relationship–specific investments are vulnerable to ex post opportunism. For example, Williamson (1975) is concerned primarily with complexity and bargaining breakdown (coordination failure) rather than the irrelevance of sunk costs in bargaining.

specifically refers to non–cooperative bargaining games. However, the solution concepts he employs all rely on backward induction. Thus, if the bargaining game has a unique equilibrium which survives backward induction, there is no way that sunk costs can affect the bargaining outcome. This is the natural starting point of our paper, which asks the following two questions:

- What if the bargaining game has multiple subgame perfect equilibria?
- What if we employ other solution concepts than subgame perfection?

The bargaining game proposed by Nash (1953), known as the Nash demand game, is a prominent example of a bargaining game with multiple equilibria. Since this game only has one stage, all the equilibria are subgame perfect. If investment is followed by the Nash demand game, is straightforward to show that efficient investment can be sustained in some subgame perfect equilibrium. Even more strikingly, if we refine the set of equilibria using the concept of stochastic stability, the only stable outcome is efficient investment. The allocation of surplus is also pinned down rather precisely, with the investor's share decreasing in the coarseness of the investment decision. If investment is binary, virtually all the surplus goes to the trading partner; as the grid of possible investments gets very fine, the investor appropriates virtually all the surplus.

Multiplicity of subgame perfect equilibria is in fact endemic in bargaining games. Other examples include the contracting game proposed by Young (1998) and alternating offers bargaining over a stream of services, as shown by Haller and Holden (1990) and Fernandez and Glazer (1991). Remarkably, van Damme, Selten and Winter (1990) demonstrate that even the celebrated alternating offers model associated with Ståhl (1972) and Rubinstein (1982) has multiple subgame perfect equilibria under the realistic assumption that the pie to be shared is not infinitely divisible.[2] For example, it may be impossible to divide money into smaller units than cents. Hence, the potential to sustain efficient investment in a subgame perfect equilibrium is not at all limited to a small or peculiar class of bargaining games.

Moreover, even if a bargaining game has a unique subgame perfect equilibrium, it is not clear that we should neglect other Nash equilibria of the game. Recent work in evolutionary game theory, notably Nöldeke and Samuelson

[2]Binmore et al. (1996) have argued that complexity costs may reestablish uniqueness of equilibrium in this game, by favoring "simple" bargaining strategies which do not depend on whether the player moves first or second at the start of the bargaining game. However, this construction becomes rather artificial once there is an additional asymmetry between the players; in our model, it is important whether the player has made an investment at the first stage or not.

(1993) and Binmore and Samuelson (1999), has made clear that evolution may not always favor subgame perfect equilibria. Evidence from bargaining experiments has also cast doubt on the predictive power of subgame perfection. To examine this issue, we also analyze a very simple bargaining game with a unique subgame perfect equilibrium, namely, an ultimatum game in which the investor can only accept or reject the proposal made by the trading partner. For believers in subgame perfection, this game provides the starkest possible example of a hold–up problem: The investor should accept any offer, no matter how meager, hence the trading partner should demand (virtually) the whole surplus, and consequently the investor should make (virtually) no investment. While this outcome is stochastically stable, so is almost any other outcome, including the efficient investment level.

Although these are two specific examples of bargaining games, the results suggest that a general principle might be at work: When there is no tension between efficiency and subgame perfection, stochastic stability not only picks some efficient outcome, but also selects a unique such outcome. On the other hand, when efficiency and subgame perfection are in conflict, stochastic stability has little cutting power. If this is indeed a general feature, the two bargaining games we study represent opposite extremes, and evolutionary analysis of the hold–up problem using other non–cooperative bargaining games is bound to admit efficient investment too.

While our analysis is evolutionary, the results can be given a rationalistic interpretation using the concept of forward induction. The reason why the investor is able to capture a share of the surplus that is sufficient to cover sunk costs is that the trading partner believes that it will not pay to be more greedy. After all, there are only two possible reasons why the investment was undertaken. Either the investor expected coordination on a favorable equilibrium or he made a mistake. The forward induction argument says that the trading partner should try to square the observed action with rational behavior; hence the trading partner should act in accordance with an equilibrium which makes the observed level of investment ex post rational for the investor. In a companion paper, Ellingsen and Johannesson (2000) presents a partial analysis of this issue.[3]

Let us briefly comment on some related literature. Young (1993b) is the first evolutionary analysis of a bargaining game. As most bargaining theory, the paper considers bargaining over an exogenously given bargaining set. Using the concept of stochastic stability, Young shows that the unique sta-

[3]Ellingsen and Johannesson also investigates other theoretical issues such as the roles of social preferences and incomplete information. They also present some experimental evidence.

ble equilibrium of the Nash demand game is the Nash bargaining solution. Subsequent evolutionary analysis of the Nash demand game, or closely related games, include Skyrms (1996), Ellingsen (1997), Young (1998) and Saez–Marti and Weibull (1999). The first evolutionary analysis of ultimatum bargaining is Gale, Binmore and Samuelson (1995). They recognize that conventional evolutionary analysis, of the kind employed in this paper, yields only the subgame perfect outcome in the ultimatum game. However, considering a particular form of drift in strategies, which gives higher probability to less costly mistakes, it becomes possible to sustain some Nash equilibria which give positive payoff to the responder.[4] In contrast to this literature, we endogenize the size of the pie over which the parties bargain, and we find that it is illegitimate to separate the analysis of the bargaining stage from the prior investment stage, as has been common not only in the hold–up literature, but also in the bargaining literature more generally. While the notion of an exogenously given bargaining set has facilitated rigorous analysis, our findings imply that if agents affect the bargaining set through prior actions, details about the prior actions need to be an integral part of bargaining theory. The evolutionarily stable bargaining outcome is affected not only by the cost and benefit associated with the actual investment, but also by the cost and benefit of the best alternative investment.

Arguably, the paper most closely related to ours is Nöldeke and Samuelson (1993), who developed the analysis of evolutionary stability in extensive form games.[5] One of their main applications is to outside option games, where the first stage is that one player selects whether to play a game with a second player or not. When there is only one feasible investment level (except no investment), our games are outside option games. Our result that there is investment in the Nash demand game is then a corollary of Nöldeke and Samuelson's Proposition 8. We go quite a bit further, both by extending the analysis to arbitrary many investment levels and by pinning down the precise sharing of surplus in a stable equilibrium of a bargaining game. However, from the perspective of game theory, our paper is primarily an application of the idea that evolution favors forward induction to a greater extent than rationalistic arguments do.

The paper is organized as follows. In Section 2, we describe the investment options and the bargaining games. The section also provides a characterization of the model's subgame perfect equilibria. In Section 3 we introduce the evolutionary process and the stability criteria. Section 4 contains the main results, and Section 5 concludes. An Appendix contains all the proofs.

---

[4]Binmore and Samuelson (1999) generalizes this idea considerably.

[5]After circulating our paper, we were informed of the work by Tröger (2000), which is by far the most closely related paper to ours. We comment on the relationship in the final section.

2 Investment and Bargaining

There are two players, A and B, who play a two–stage game. At stage 1, player A chooses an investment I from a finite set $\Psi = \{0 = I_0, I_1, ..., I_N = \bar{I}\}$. This investment creates a benefit (pie) of size $V(I)$. At stage 2, the players bargain. We will consider two different bargaining games, namely, the Nash demand game and the ultimatum game. In either case both players demand a fraction of the pie. Let $D(I) = \{\delta, 2\delta, ..., V(I) - \delta\}$ denote the set of possible demands that a player can make.[6] Player A's demand is denoted y and player B's demand is denoted x. For simplicity, we assume that $V(I)/2$ is divisible by δ.

The pie, $V(\cdot)$, is strictly increasing in investment. Let $I^* = \arg\max\{V(I) - I | I \in \Psi\}$ denote first–best investment, and let $\hat{I} = \arg\max\{V(I) - I | I \in \Psi \backslash I^*\}$ denote the investment level which generates the second highest total utility. Both I^* and $\hat{I}$ are assumed to be unique.[7]

Bargaining according to the *Nash demand game* proceeds as follows. After observing I, both players simultaneously make their demands y and x. If $x + y \leq V(I)$ each player gets a share of the pie equal to his demand. Otherwise, each player gets nothing. Taking investment costs into account, the payoff of player A can thus be denoted

$$\pi_A = \begin{cases} y - I, & \text{if } x + y \leq V(I); \\ -I, & \text{otherwise.} \end{cases}$$

Player B's payoff is

$$\pi_B = \begin{cases} x, & \text{if } x + y \leq V(I); \\ 0, & \text{otherwise.} \end{cases}$$

Observe that for player A, a pure strategy for the whole game is a pair $(I, y) \in \Psi \times D(I)$, whereas a pure strategy for player B is a function $x : \Psi \to D(I)$.

Alternatively, bargaining proceeds according to the rules of the *ultimatum game.* In this case, player B makes an offer x which player A observes and then either accepts or rejects. If the offer is accepted, player B gets x and player A gets $V(I) - x$. Otherwise, both players get nothing. Note that the payoffs are the same as with the Nash demand game, but that A's set of demands

[6] Finiteness of the set D can be justified by the fact that there is often a smallest unit of account, whereas finiteness of Ψ can be given the additional justification of technological indivisibility. Above all, finiteness of D and Ψ facilitates the evolutionary analysis, because it allows the space of relevant strategies to be finite.

[7] If there are two optimal investment levels, then choice between these two can work like cheap talk, which would trivialize the analysis. It would make no difference if there were two levels of investment which tied for second best.

is smaller; it contains only the elements $D_A = \{V(I) - x, V(I)\}$, where the first element is equivalent to accepting B's offer and the second element is equivalent to rejecting it. However, the smaller set of demands for A is not an important difference between the two games. It would not matter much if we allowed player A to choose any demand in D; the essential distinction between the two bargaining games is that B has a first–mover advantage in the ultimatum game. For player A a pure strategy for the whole game is a now pair $(I, y(x))$, i.e., A's demand is a function $y : D \to D$ rather than a real number.

Before turning to the evolutionary analysis, let us consider the subgame perfect Nash equilibria. When the investment decision is followed by the Nash demand game, there is a vast multiplicity of subgame perfect equilibria. In particular, there are subgame perfect equilibria sustaining first–best investment. To take one example, suppose A plays the strategy $(I = I^*, y = V(I^*) - \delta)$, and B plays the strategy $x(I) = \delta$ for all I. Clearly, B plays a best response to A's strategy, since any higher demand than δ leads to a payoff of zero. To see that A also plays a best response, notice that A by claiming $V(I) - \delta$ is a residual claimant to the surplus from investment. Thus first–best investment can be implemented in a subgame perfect equilibrium without any contract whatsoever. This observation is in stark contrast to the conventional wisdom in the hold–up literature, which is that investment will be distorted if the division of surplus can not be committed to ex ante. That conclusion is inevitable in the hold–up literature because it assumes that the bargaining game has a unique subgame perfect equilibrium. In that case, players always get a fixed share of the pie, and there tends to be underinvestment. Once there are many subgame perfect equilibria, as above, the possibility emerges that the private marginal return to investment can be the same as the marginal social return. On the other hand, there are also equilibria with no investment. For example, it is an equilibrium of the bargaining stage that B demands $V(I)$ and A demands 0, and this equilibrium in the subgame obviously sustains $I = 0$.

If the investment decision is followed by the ultimatum game, any subgame perfect equilibrium outcome entails a low level of investment, $I \leq \delta$. In this case, player B knows that at the bargaining stage A will accept any proposal which yields a positive share of the pie. Hence, player A gets at most δ from the bargaining stage. Thus, it would be irrational for A to invest more than δ.

The above arguments extend straightforwardly, allowing a simple characterization of subgame perfect equilibrium outcomes for the two games.

Proposition 1 *(i) If surplus is divided according to the Nash demand game, any investment level $I \in \Psi$ can be sustained in a subgame perfect equilibrium. (ii) If surplus is divided according to the ultimatum game, only investment levels $I \leq \delta$ can be sustained in a subgame perfect equilibrium.*

Subgame perfection admits virtually any outcome in the Nash demand game, but only admits low investment in the ultimatum game. As shown below, the conclusion is radically different when we apply the criterion of evolutionary stability instead of subgame perfection.

3 Evolution

Evolutionary analysis refrains from answering the metaphysical question of which strategies are rational. Instead, it tries to answer the more economic question of which strategies can be expected to survive competitive pressures. The study of stochastic evolution in games was pioneered by Foster and Young (1990), Kandori, Mailath and Rob (1993) and Young (1993a). The extension to extensive form games is due Nöldeke and Samuelson (1993), and we apply their framework here.

For each player role, A and B, let there be a *population* of size N. Each period $t \in \{1, 2, ...\}$ every possible combination of agents in populations A and B meet and play the investment cum bargaining game. The set of strategies is the same as above. Agents also hold *beliefs* about their opponent, but taking an evolutionary approach, we do not require players to behave rationally given their beliefs. Let $\nu(\cdot|I)$ denote player A's beliefs concerning player B's demand, and let $\sigma(\cdot|I)$ denote player, B's belief about player A's demand. Both ν and σ are probability distributions on the set of possible demands, and they are contingent on the investment I. If surplus is divided according to the ultimatum game, σ also depends on player B's demand, x.

We make two additional "technical" assumptions.

Assumption 1 *(i) The pie division is small:* $V(\hat{I}) > 2\delta$. *(ii) The population is large:* $V(I^*)/N < \delta$.

The first assumption assures that the pie division is small enough that the investor can actually get a positive net surplus with the second best investment. The second assumption assures that the population is large enough (relative to the minimum division) so that if one agent in the other population changes his demand this will not change the best response demand.

A *state* θ specifies how many agents in each population have each possible combination of belief and strategy. Observe that the set of possible states, denoted Θ, is finite. (Since the space of strategies and beliefs is different for the two games, so is the state space.) With each state θ there is an associated probability distribution of terminal nodes, denoted $z(\theta)$.

Beliefs and strategies evolve in two different ways; by *adaptation* to the current environment and by *random mutation.* Adaptation occurs in the following way. Every period each agent has an i.i.d. chance of rationally updating his beliefs and strategy. This is called an updating draw. An updating agent observes $z(\theta)$, updates his beliefs based on this observation (beliefs following decision nodes not reached in state θ are unchanged) and chooses a best response to his new beliefs. Updating works on behavioral strategies. If an agent is already playing a best (behavioral) response following some node, then he continues to do so, if not then he chooses one of the available best responses, each with positive probability. Agents' beliefs and strategies are also changed by 'mutation.' Every period each agent has an i.i.d. probability ϵ of mutating. When an agent mutates his beliefs and strategy are chosen at random from some exogenously given distribution which gives full support to all of that agent's possible belief/strategy combinations. The updating draw and mutations combine to form a markov chain over the state space Θ in which every transition has positive probability. Hence there exists an ergodic distribution $\mu(\epsilon)$.

If a mutation changes an agent's action only at a decision node that is not currently reached, or only changes an agent's beliefs following an unreached node, then this has no effect on the agent's payoff. The process of moving a population's actions or beliefs in this way is called *drift*.

3.1 Solution concepts

We are now ready to introduce our main solution concepts, *stochastic stability* and *local stability.* The set of stochastically stable states, denoted Θ^*, are those assigned positive probability in the limit distribution $\mu^* = \lim_{\epsilon \to 0} \mu(\epsilon)$. Local stability is weaker than stochastic stability. A set of states is locally stable if it takes more than one mutation to escape that set. An even weaker solution concept is that of an *absorbing set.* A set Q is absorbing (w.r.t. the updating draw) if the following two conditions hold: (i) from a state $\theta \in Q$ it is impossible to get to any state outside of Q without mutation, and (ii) if θ and θ' are both elements of Q, then it is possible for the population to get from state θ to θ' without mutation. Let $\bar{Q}$ be the set of absorbing sets. If an absorbing set Q has only one element, θ, then we call θ an *equilibrium.*[8] As it turns out, all absorbing sets are equilibria in our model.

Proposition 2 *Under either bargaining rule, if Q is an absorbing set, then Q is a singleton.*

Proposition 2 allows us to speak about equilibria rather than absorbing sets from now on. Let $\bar{\Theta}$ be the set of equilibria.

[8]Observe that an equilibrium thus defined need not be a Nash equilibrium; rather, it is a self–confirming equilibrium in the sense of Fudenberg and Levine (1993).

Before getting to the more substantial results, we need to be more precise about our definition of local stability. The *basin of attraction* of an equilibrium θ, denoted $B(\theta)$, is the set of states θ' such that the population can get from θ' to θ without mutation. Similarly, we say that θ' is in the *single mutation neighborhood* of θ, denoted $\theta' \in M(\theta)$, if θ' and θ differ by a single mutation. A union of equilibria, X, is a *mutation connected set* if for all pairs of equilibria θ_1,$\theta_n \subset X$, there exists some ordering of the remaining equilibria, $(\theta_2, ..., \theta_{n-1})$, such that for all $k = 1, ..., n$, $M(\theta_k) \cap B(\theta_{k+1}) \neq \emptyset$. I.e., the set is mutation connected if one can get from one equilibrium to another through a sequence of single mutation transitions. A set of equilibria X is locally stable if: (i) X is mutation connected, and (ii) from any $\theta \in X$, a single mutation (followed only by adaptation) is not enough to exit from X. (Formally, the latter requirement is that if $\theta \in X$ and $\theta' \notin X$, then $M(\theta) \cap B(\theta') = \emptyset$.)

As mentioned above, local stability is a necessary condition for stochastic stability. Moreover, it can be shown that if one state in a locally stable set is stochastically stable, then all states in that set are stochastically stable (Nöldeke and Samuelson, 1993, Proposition 1). Let Θ^L denote the union of all locally stable sets.

Since the same equilibrium outcome (I, y, x) can often result from many different states (because beliefs may differ), it is sometimes useful to speak of the stability of outcomes rather than the stability of states. Likewise we sometimes speak of transitions from one outcome to another as shorthand for "transition from an equilibrium with one outcome to an equilibrium with another outcome." An outcome ρ is called locally stable if the set $\{\theta | z(\theta) \text{puts probability 1 on } \rho\}$ is a locally stable set. It will be shown below that every locally stable set corresponds to a locally stable outcome, so that there is no loss of generality in focussing on outcomes.

4 Main results

Our first result is a characterization of the locally stable outcomes when the parties bargain according to the Nash demand game. As it turns out, only first–best investment is locally stable. Any other investment level can be escaped from by a simple series of single mutation transitions. To see this, suppose the population has settled on an equilibrium involving a suboptimal investment level I. Let population B drift to believe with certainty that were A to invest I^*, then the associated demand would be $V(I^*) - \delta$. Then, if a member of population A mutates to investing I^* and claiming $V(I^*) - \delta$, this agent will do better than all other agents in population A. Hence, as other agents in population A update their strategy, they will also start playing $(I^*, V(I^*) - \delta)$.

While all locally stable outcomes have investment I^*, there is some scope for variation in the equilibrium division of surplus. We show that the largest demand by agent B which is consistent with local stability is

$$x^L = \max\{x \in D_B(I^*) | (V(I^*) - x)\frac{N-1}{N} - I^* \geq V(\hat{I}) - \hat{I} - \delta\}.$$

To understand the magnitude of x^L, observe that the largest demand that agent B could make following an efficient investment, that would not give agent A an incentive (weakly) to choose a less efficient investment, even if he expected to get (almost) all of the surplus is

$$x^M = \max\{x \in D_B(I^*) | V(I^*) - x - I^* > V(\hat{I}) - \delta \ - \hat{I}\}.$$

It is straightforward to show that $x^M - \delta \leq x^L \leq x^M$. The steps of the argument are the following:

$$\begin{aligned}(V^* - x^M + \delta)(N-1)/N - I^* &= V(I^*) - x^M - I^* - (V(I^*) - x^M)/N \\ &\quad +\delta(N-1)/N \\ &> V(I^*) - x^M - I^* - (\delta - x^M)/N \\ &\geq V(I^*) - x^M - I^* \\ &\geq V(\hat{I}) - \delta - \hat{I},\end{aligned}$$

where the strict inequality is due to Assumption 1. Intuitively, the $(N-1)/N$ term in x^L assures that if one agent in population B changes his demand, this will not cause agents in population A to change their investment away from the efficient level.

Proposition 3 *Let agents bargain according to the Nash demand game. The outcome ρ is locally stable if and only if $\rho = \{(I^*, V(I^*) - x, x)\}$, where $x \leq x^L$.*

Note how, in this case, local stability identifies a much smaller set of outcomes than did subgame perfection. In particular, subgame perfection allowed inefficient investment, whereas local stability does not.

Let us now refine the set of locally stable outcomes and consider the smaller set of stochastically stable outcomes. Although stochastic stability is easy enough to define, the computation of stochastically stable equilibria is a bit more demanding. It basically requires counting the number of mutations needed to move from one equilibrium to another. The equilibria which are most easily reached from all other equilibria (in terms of requiring fewest mutations) are stochastically stable. To articulate this idea precisely, we need a couple of additional definitions. Recall that $\bar{\Theta}$ denotes the set of equilibria. Let $r(\theta, \theta')$ be the minimum number of mutations needed to move from an equilibrium θ to another equilibrium, θ'. Define the graph $\mathcal{G}$ as the collection of vertices, one vertex for each equilibrium, with a directed edge from every vertex to every other. The *resistance* (or *cost*) of the edge $\theta \rightarrow \theta'$ is $r(\theta, \theta')$. A *θ–tree,* Γ, is a collection of edges in $\mathcal{G}$ such that from every vertex $\theta' \neq \theta$ there is a unique directed path to θ, and there are no cycles. The resistance of a tree Γ is the sum of the resistances of all the edges in the tree. Finally, the *stochastic potential* of an equilibrium θ is the minimum resistance over all θ-trees. The key to checking whether an equilibrium is stochastically stable is provided by Young (1993a, Theorem 4).

Theorem 1 *An equilibrium θ is stochastically stable if and only if no other equilibrium has lower stochastic potential.*

In fact, it is shown below (in the proof of Proposition 4) that it suffices to construct trees that are much simpler than those described above. By definition, any transition between equilibria requires at least one mutation. Hence, when constructing a minimum resistance tree, one can ignore edges with resistance 1. Notice also that from any equilibrium one may arrive at an equilibrium within a locally stable set through a sequence of one mutation transitions, and one may move around within the locally stable set in the same manner. Hence, it suffices to construct trees with locally stable sets (represented by locally stable outcomes) as vertices.

Since all locally stable outcomes under Nash demand bargaining have the property that $I = I^*$ and that $y = V(I^*) - x$, these outcomes can be fully characterized by B's demand, x. There are two ways in which a transition between locally stable outcomes may occur. First, there might be a direct transition, during which investment is maintained at the efficient level. In this case, we can appeal to Young (1993b) and observe that if $x < x^{NBS} = V(I^*)/2$, then the easiest transition is to the outcome $x + \delta$. Conversely, if $x > x^{NBS}$, the easiest transition is to $x - \delta$. Denote the resistance to such a transition by $r(x)$. Second, there might be an indirect transition, during which the population passes through a state with inefficient investment. This involves having a sufficiently large portion of population B increase their demands to such an extent that efficient investment is less attractive to population A than the outcome $(\hat{I}, V(\hat{I}) - \delta, \delta)$ (which is the most attractive inefficient outcome). If this happens—as it might do, following appropriate drift—then the populations will make a transition to this inefficient outcome, after which a sequence of single mutation transitions (the resistances of which we can ignore) suffice to get the populations to $(I^*, V(I^*) - \delta, \delta)$. Let $\hat{r}(x)$ denote the resistance of the transition from the outcome $(I^*, V(I^*) - x, x)$ to $(\hat{I}, V(\hat{I}) - \delta, \delta)$. Since we can show that $\hat{r}(x) > r(x)$ whenever $x > x^L$ (see Appendix), it is easy to construct a minimum resistance tree. The case $x^M > x^{NBS}$ essentially reduces to the analysis of the Nash demand game in Young (1993b). Otherwise, if $\hat{r}(x^L) \geq r(x^L - \delta)$, the minimum resistance tree is given by

$$\delta \longrightarrow 2\delta \longrightarrow ... \longrightarrow x^L - \delta \longrightarrow x^L,$$

while if $\hat{r}(x^L) \geq r(x^L - \delta)$, the minimum resistance tree is given by

$$x^L \longrightarrow \delta \longrightarrow 2\delta \longrightarrow ... \longrightarrow x^L - \delta.$$

Both trees are constructed by including the minimum resistance transition out of each locally stable outcome and then deleting the transition among these which has the highest resistance.

Proposition 4 *Let surplus be divided by the Nash demand game.*

(i) If $x^{NBS} < x^M$, then the unique stochastically stable outcome is $(I^*, V(I^*) - x^{NBS}, x^{NBS})$.

(ii) If $x^{NBS} \geq x^M$, then the set of stochastically stable outcomes is contained within $\{(I^*, V(I^*) - x, x) | x \in \{x^L - \delta, x^L\}\}$.

For a more precise statement of part (ii) of the proposition, see Lemma 10 in the Appendix. From the point of view of the hold–up problem, part (i) is quite uninteresting; the difference between efficient and inefficient investment is so large that even the Nash bargaining solution would yield a sufficient investment incentive.

Proposition 4 implies that if the set of possible investments represents a sufficiently fine grid, almost all the surplus goes to the investor, agent A. To see this, note that x^M is the maximum demand such that it would not pay for the investor to invest the second–best level, $\hat{I}$, even if he could keep virtually all the surplus. Hence, if $\hat{I}$ is close to I^*, the investor must get almost all the surplus from I^* as well.

If, on the other hand, the difference between first–best and second–best investment is large, then a substantial fraction of the surplus may go to the trading partner, agent B. In fact, if investment is binary, virtually all the surplus may go to agent B. For example, suppose investment is either 0 or 60, and that $V(0) = 0$ and $V(60) = 100$. Then $x^{NBS} = 50 > x^M = 40$. Since $\hat{r}(40) < r(40 - \delta)$, the unique stochastically stable equilibrium outcome is $(I^*, 60 + \delta, 40 - \delta)$, leaving the investor with a minimal profit. An interpretation of this result is that it goes as close as possible to the ex post equal split while respecting the forward induction property, identified by Nöldeke and Samuelson (1993), that if there is any strict equilibrium consistent with investment, then there must be investment in any stable equilibrium.[9]

[9] Interestingly, an experiment by Binmore et al. (1998) concerns what is essentially the

Suppose now that bargaining is conducted according to the rules of the ultimatum game instead. Let $x^{\max}(I)$ be the largest demand which is weakly smaller than $V(I)$, and let I^H be such that $V(I^H) - I^H - x^{\max}(I^H) > V(I) - I - x^{\max}(I)$ for all $I \neq I^H$. In other words, I^H is the investment level if agent A expects to be held up.

Proposition 5 *Let surplus be divided according to the ultimatum game. An equilibrium is stochastically stable if and only if agents in population A receive at least $V^H - I^H - x^{\max}(I^H)$.*

Thus, stochastic stability has no cutting power when agents bargain according to the ultimatum game. Although there is a unique subgame perfect equilibrium, with no investment, other equilibria are equally stable. In particular, first–best investment can be sustained in a stable equilibrium. Likewise, there is no prediction about how the surplus generated by investment will be shared between the two agents; any sharing of the net surplus is stable.

To understand why the set of stochastically stable equilibria is so large under ultimatum bargaining, note for example how easily populations can go from the least efficient equilibrium to the most efficient equilibrium. Let population B drift to believe that if an agent from population A makes the investment I^*, then that agent will reject demands which leave him with (weakly) less than he got in the least efficient equilibrium. Then, a single mutation by a member of population A, to the investment level I^*, is enough to move the whole population to the most efficient equilibrium in the next round. Conversely, at an efficient equilibrium, population A may drift to accept greedier offers, whereupon it takes only a single mutation to an agent in population B to destabilize the equilibrium.

5 Final remarks

The idea that current bargaining outcomes are independent of sunk costs has been applied widely both in economics and in other social sciences. In the recent literature on the hold–up problem, Hackett (1994) is a lone dissenting voice, pointing to sociological research and to his own experiments for evidence that sunk costs matter. Hopefully, this paper has shown that game theory admits a role for sunk costs as well.

An obvious limitation of our analysis is that we have considered two specific bargaining games. A useful extension of the present paper would be to

---

normal form of this game. Using a different argument, they argue that the outcome which gives all the surplus to the trading partner is natural if players are (close enough to) being rational.

replicate our analysis for other bargaining games. We conjecture that when investment is followed by a relatively symmetric bargaining game, the results will be quite similar to our results for the Nash demand game. As indicated in the introduction, this might be true even for the alternating offer bargaining model of Rubinstein (1982).

Our analysis of the hold–up problem has focussed on the extreme case that parties can write no ex ante contract whatsoever. But the paper's message is also relevant when written contracts are subject to renegotiation. A vast recent literature has considered incomplete contracts which partially or fully solve the hold–up problem; see in particular Maskin and Tirole (1999) and the references therein. One response to this literature has been given by Hart and Moore (1999), who argue that the the hold–up problem tends to reappear when parties cannot contract on the renegotiation procedure. Hart and Moore instead assume that surplus from renegotiation is split with no regard to investment costs. The current paper suggests that this bargaining assumption is questionable.

After circulating our paper, we were informed of the independent and closely related work of Tröger (2000). There is substantial overlap between the two papers. In particular, Tröger proves the equivalent of our Proposition 4. Besides easily identifiable differences in packaging, we would like to point out two distinguishing features. First, the evolutionary dynamics are not the same. Tröger extends the model of Young (1993a), so as to make it applicable to extensive form games. In contrast, we have applied the extensive form framework of Nöldeke and Samuelson (1993) (which in turn builds on Kandori, Mailath and Rob (1993)). Whereas we cannot claim any methodological contribution, we think our approach offers some advantages. One advantage is that beliefs are explicitly modeled, which ties the analysis closer to conventional models of human decision making. Another advantage, which is only apparent ex post, is that Nöldeke and Samuelson's framework enables simpler analysis of the problem at hand. Our proofs are both shorter and more elementary than Tröger's. At the same time, it is assuring that the central results are insensitive to the details of the evolutionary process. In this sense, the two papers are mutually supportive. A second difference between the papers is that we present results for ultimatum bargaining as well as for the demand game.

Appendix: Proofs

We start by proving Propositions 2 and 3. Let $\rho(Q)$ denote the set of outcomes associated with the absorbing set Q.

The following two lemmas are needed in order to prove Proposition 2.

Lemma 1 *Let $z_1 < z_2 ... < z_k$ be demands in $D(I)$ for some $I \in \Psi$. Assume that the set of demands following I for agents in the relevant population is $\{z_l\}_{l=1}^k$. Then the set of best (behavioral) response demands following I for agents in the other population is a subset of $\{V(I) - z_l\}_{l=1}^k$.*

Proof: Let N_l be the number of agents making demands of z_l or less. The payoff (ignoring the cost of investment $-I$ for population A which is held constant) to a demand of w with $V(I) - z_{l+1} < w < V(I) - z_l$ is $wN_l/N < N_l(V(I) - z_l)/N$ (which is the payoff for a demand of $V(I) - z_l$). The payoff for a demand $w > V(I) - z_1$ is $0 < N_1(V(I) - z_1)/N$ (which is the payoff for a demand of $V(I) - z_1$). The payoff for a demand $w < V(I) - z_k$ is $w < V(I) - z_k$ (which is the payoff for a demand of $V(I) - z_k$). Hence any demand not in $\{V(I) - z_l\}$ can be improved upon by a demand which is in that set. $\square$

Lemma 2 *If Q is an absorbing set with $\{(I, y, x), (I, y', x')\} \in \rho(Q)$ and $x \neq x'$ or $y \neq y'$, then Q is a singleton.*

Proof: From Lemma 1 we know the set of demands following I must be $\{x_l\}$ for population B and $\{y_l = V(I) - x_l\}$ for population A. Let N_l be the number of agents making demand x_l following I, and let M_l be the number of agents investing I and then demanding y_l. We assert that these numbers must be constant in Q. Otherwise let a single agent i update and change N_l (resp. M_l.) Let $k \neq l$, clearly it can not be the case that both y_l and y_k (resp. x_l and x_k) are both best responses following I both before and after agent i has updated. Start from this state in which one of these demands is not a best (behavioral) response following I, and let all agents making that demand update. That demand (call it z_l) is now no longer made following I. Let all agents in the other population update. They now have beliefs (by Lemma 1) following I such that they will not make demand $V(I) - z_l$ following I unless they observe a demand of z_l following I. However, in the next state in which investment I is made, let all agents in the first population update, they now believe that demand $V(I) - z_l$ is not made, and consequently will not make demand z_l. Hence these two demands have disappeared and can not reappear, which contradicts the assumption that Q is an absorbing set. Since M_l and N_l can't decrease, (I, y_l) is a best response and x_l is a best (behavioral) response following I. Since they can't increase, every other strategy being played must do as well, therefore the population is in equilibrium and Q is a singleton. $\square$

We are then ready to prove Proposition 2.

Proof of Proposition 2: Assume that Q is not a singleton. Since updating does not change off–path beliefs, $\rho(Q)$ must be a nonsingleton. Consider $\theta \in Q$ in which investment I' is made by some agent in population A. By Lemmas 1 and 2, we know that if Q is not a singleton, then $\forall (I, y, x) \in \rho(Q)$, $x + y = V(I)$. Let $(I, y, x) \in \rho(Q)$ be such that if $(I', y', x') \in \rho(Q)$, then $y - I \geq y' - I'$. If $y - I > y' - I'$, then agents playing (I', y') will update to (I, y) (or a strategy which does equally well), but agents would never update to (I', y') as long as someone is playing (I, y). Hence it must be the case that $y - I = y' - I'$, or else Q was not an absorbing set. However, if $y - I = y' - I' \; \forall (I', y', x') \in \rho(Q)$, then agents will update away from neither (I, y), nor (I', y'). Thus, updating alone can not change the state, and Q is a singleton. $\Box$.

The next two lemmas are used in the proof of Proposition 3. The first ensures uniqueness of equilibrium outcomes.

Lemma 3 *Let θ be an equilibrium such that $\rho(\theta)$ is not a singleton. Then $\exists \theta'$, an equilibrium, such that $\rho(\theta')$ is a singleton and such that the population can get from θ to θ' through a sequence of single mutation transitions.*
If in addition, every investment in θ is followed by a unique pair of demands, and $(I, y, x) \in \rho(\theta)$, then θ' can be chosen so that $\rho(\theta') = \{(I, y, x)\}$.

Proof: Since θ is an equilibrium, it must be the case that all agents in population A receive the same payoff. Consider first the case where multiple demands are made following some investment I'. Let the demands made by population B following I' be $x_1 < ... < x_k$, and let a single agent who demands x_k following I' mutate to demand x_1 following I'. In the following period let all agents in population A update, they will all choose $(I', V(I') - x_1)$ which is now their unique best response. In the following period let all agents in population B update, they all switch to a demand of x_1 following I'. We are now at the desired equilibrium.

Now assume a single demand is made following each investment level. Let a single agent playing (I', y') $(I' \neq I)$ mutate to (I, y). Since the distribution of demands following any investment has not changed, no agent in population B has an incentive to change strategy. No new information has been revealed (there were already agents playing (I, y),) and so no agent in population A has an incentive to change his strategy. Therefore we are at a new equilibrium with one more agent playing (I, y). By repeating this process we arrive at the equilibrium θ' in the Lemma. $\Box$

Lemma 4 *Let θ' $(\rho(\theta') = \{(I', y', x')\})$ be an equilibrium. If $I \neq I'$ and $y - I \geq y' - I'$, then the population can get from θ' to an equilibrium θ with $\rho(\theta) = \{(I, y, x)\}$ through a sequence of single mutation transitions.*

Proof: In state θ' let agents in population B drift to believe with certainty that any agent in population A that invests I will demand y. This implies that for all j in population B, $x_j(I) = x = V(I) - y$. Let a single agent in population A mutate to play (I, y) (no change to beliefs.)[10] In the next period let all agents in A update, they all observe that $x_j(I) = x$ for all j in population B. If $y - I > y' - I'$ then their best response is (I, y) to which they all switch, leaving them at an equilibrium θ with $\rho(\theta) = \{(I, x, y)\}$. If $y - I = y' - I'$ then all agents in population A are playing a best response and we are at a new equilibrium θ_1 with $\rho(\theta_1) = \{(I, y, x), (I', y', x')\}$. An application of Lemma 3 gets us to a state θ with $\rho(\theta) = \{(I, y, x)\}$. □

Proof of Proposition 3: We need to demonstrate (i) that the population can get from any equilibrium with an outcome not satisfying the Proposition's characterization to an equilibrium with an outcome which does satisfy it, through a sequence of single mutation transitions, and (ii) that the population can not depart from an outcome satisfying the characterization without at least two simultaneous mutations. Step (i): by Lemma 3 we may consider only θ such that $\rho(\theta)$ is a singleton. Let $\rho(\theta) = \{(I, V(I) - x, x)\}$. If $I \neq I^*$ then $V(I^*) - \delta - I^* > V(I) - x - I$, so by Lemma 4 the population can get to an equilibrium θ^L with $\rho(\theta^L) = \{(I^*, V^* - \delta, \delta)\}$. If $I = I^*$ but $x > x^L$ and $V(\hat{I}) - \delta - \hat{I} \geq V(I) - x - I$, then by Lemma 4 the population can get to an equilibrium $\hat{\theta}$ with $\rho(\hat{\theta}) = \{(\hat{I}, V(\hat{I}) - \delta, \delta)\}$. Another application of Lemma 4 then gets the population to an equilibrium θ^L with $\rho(\theta^L) = \{(I^*, V^* - \delta, \delta)\}$. Finally, it might be that $I = I^*$, $x > x^L$ and $V(\hat{I}) - \delta - \hat{I} < V(I) - x - I$, but $V(\hat{I}) - \delta - \hat{I} > (V(I) - x)(N-1)/N - I$. In this case, allow agents in population B (resp. A) to drift to believe with certainty that population A (resp. B) agents will demand $V(\hat{I}) - \delta$ (resp. δ) following an investment of $\hat{I}$. If a single agent in population B mutates to demand $x + \delta$ following demand of I^*, then population A payoffs will drop below what they correctly expect to get following an investment of $\hat{I}$, which will lead them to play (following updating) $(\hat{I}, V(\hat{I}) - \delta)$. As above, an application of Lemma 4 completes the proof.

Step (ii): Consider some θ with $\rho(\theta) = \{(I^*, y, x)\}$, and $x \leq x^L$. We must show that a single mutation can only move the population to a state θ_1 with $\rho(\theta_1) = \rho(\theta)$. Note first that for $I \neq I^*$, $V(I) - \delta - I \leq V(\hat{I}) - \delta - \hat{I} < (V(I^*) - x^L)(N-1)/N - I^* \leq (V(I^*) - x)(N-1)/N - I^*$. Hence, agents in Population A will never choose an investment level other than I^* as long as they believe that at most one agent in population B is demanding more

[10]Note that since beliefs are not updated following mutation we might as simply say that he believes that for all j in population B, $x_j(\tilde{I}) = V(\tilde{I}) - \delta$ if $\tilde{I} \neq I$ but $x_j(I) = x$.

than x following an investment of I^*. Hence, if an agent were to make an investment other than I^* because of a mutation, updating would not cause other agents to imitate him, and he would therefore eventually update back to (I^*, y). Now consider a mutation which leaves investment unchanged, but changes demand following I^* (this may be a mutation to an agent in either population.) Since at least $(N-1)$ agents in population B still demand x following I^*, investment remains unchanged. For concreteness, let an agent in population B change his demand to $\tilde{x}$. Since $x = V(I^*) - y$, it is a best response, and so this different demand does not give other agents in population B reason to change their demand. By Lemma 1, either y or $\tilde{y} = V(I^*) - \tilde{x}$ is the best response for agents in population A. If $\tilde{x} > x$, then demanding y pays $((N-1)y + 0)/N - I^*$, while demanding $\tilde{y}$ pays $\tilde{y} - I^*$. Agents in population A will not change their demand if $\frac{N-1}{N}y \geq \tilde{y}$. From the assumption that $V(I^*)/N < \delta$, we have $y/N < \delta$ which implies that $\frac{N-1}{N}y > y - \delta \geq \tilde{y}$. If $\tilde{x} < x$ then the payoff for demanding y is $y - I^*$, while the payoff for demanding $\tilde{y}$ is $\tilde{y}/N - I^*$. Since $y > V(\hat{I}) - \delta > \delta > V(I^*)/N > \tilde{y}/N$, playing y is again the best response for population A following this mutation. The case in which an agent in population A mutates to a different demand is symmetric, except for the absence of the $-I^*$ term. Hence no single mutation causes any agent other than the mutating agent to change his behavior, and the population must return to an equilibrium θ' with $\rho(\theta') = \rho(\theta)$ as soon as the mutating agent receives the updating draw. $\square$

From now on, write V^* and $\hat{V}$ as shorthand for $V(I^*)$ and $V(\hat{I})$.

The following Lemmas are for figuring stochastic stability. Recall that we have argued in the body above that it suffices to construct trees with vertices made up of locally stable outcomes. Hence, the first step—taken in Lemmas 5, 6 and 8—is to find, for each locally stable outcome, the easiest transition from an equilibrium with that outcome to an equilibrium with another outcome. It is shown that for (almost all) locally stable outcomes, the easiest transition is one in which the investment level is not changed, and demands are changed only minimally (by $+/-\delta$). To show this, we must first find the least number of mutations for a transition from a locally stable outcome to an outcome with inefficient investment (for comparison). Now, clearly, an agent in population A will only change his investment if he thinks that he is going to get something better, and the best outcome that he could expect with an inefficient investment is $(\hat{I}, \hat{V} - \delta, \delta)$. Hence, the question becomes: how many agents in population B have to mutate to a higher demand to make the above outcome better than maintaining efficient investment?

Recall that $x^L \in \{x^M, x^M - \delta\}$.

Lemma 5 *The number of mutations required to get from an equilibrium with* *outcome (I^*, y, x) with $x \leq x^L$ to an equilibrium with outcome $(\hat{I}, \hat{V} - \delta, \delta)$ is $\hat{r}(x) = \min\{r | r > N(1 - \frac{\hat{V}-\delta-\hat{I}+I^*}{V^*-x})\}$.*

Proof: To make agents change investment, r (the number of B agents who mutate to a higher demand) must be large enough so that

$$\frac{N-r}{N}(V^* - x) - I^* < \hat{V} - \hat{I} - \delta, \tag{1}$$

since updating agents change their actions only when they are not already playing a best response. Solving for r yields the desired expression. □

Lemma 6

(i) If θ is an equilibrium with outcome (I^*, y, x), and $x < \min\{x^M, x^{NBS}\}$, the easiest transition away from $\rho(\theta)$ is to an equilibrium with outcome $(I^*, y - \delta, x + \delta)$.

(ii) If $x^{NBS} < x < x^M$, then the easiest transition away from $\rho(\theta)$ is to an equilibrium with outcome $(I^*, y + \delta, x - \delta)$.

(iii) If $x = x^M < x^{NBS}$ then an easiest transition is to an equilibrium with outcome $(\hat{I}, \hat{V} - \delta, \delta)$.

Proof: If $x < x^M$ then $V^* - x - \delta - I^* \geq V^* - x^M - I^* > \hat{V} - \delta - \hat{I}$ which implies that $V^* - x - \delta > \hat{V} - \delta - \hat{I} + I^*$. Hence

$$N(1 - \frac{V^* - x - \delta}{V^* - \delta}) < N(1 - \frac{\hat{V} - \delta - \hat{I} + I^*}{V^* - \delta}). \tag{2}$$

And so long as the population is sufficiently large, so too is the transition cost.

Now we know that if $x < x^{NBS}$ then it is easier to make a transition to demands of $x + \delta$, whereas if $x > x^{NBS}$ it is easier to make transitions to $x - \delta$, which combined with the fact that transitions to $x + \delta$ are easier than transitions to inefficient investments completes the case where $x < x^M$.

From Proposition 3, we do not need to worry about $x > x^L$, so all we need to check is the case where $x = x^L = x^M$. If $x^M < x^{NBS}$, then we know that $r(x) \geq \hat{r}(x)$, because $V^* - x^M - \delta \leq \hat{V} - \delta - \hat{I}$.□

If $x = x^M > x^{NBS}$ then we can not say any more than that the easiest transition is to an equilibrium with an outcome of either $(\hat{I}, \hat{V} - \delta, \delta)$ or $(I^*, y + \delta, x - \delta)$. We already knew this, but since either one of these transitions gets us easily to another locally stable set, and allows an easy construction of a tree around the Nash bargaining solution and efficient investment, it really does not matter.

Next, we turn to the analysis of stochastic stability. We first consider the number of mutations required to make a transition directly from an equilibrium with outcome (I^*, y, x), $(y = V^* - x)$ to one with outcome (I^*, y', x'). Along this transition we will not allow the level of investment to change for any agent. Hence results in this section are essentially borrowed from Young's bargaining paper. Later we will worry about multi–step transitions in which one first changes the investment and then changes the demand following the efficient investment.

Lemma 7 *From an outcome (I^*, y, x) the easiest transition in which investment is at all times efficient, but which ends with different demands, is to an outcome (I^*, y', x') where $x = x - \delta, x + \delta, \delta,$ or $V^* - \delta$.*

Proof: From Young (1993b) Lemma 1. □

The idea is that if one population changes their demand, then the increase in demand which will be least hard to get the other population to accept is an increase of δ, while the decrease in demand that is easiest to get the other population to accept is a decrease all the way to δ.

Lemma 8

(i) Moving from x to $x - \delta$ takes $N(1 - \frac{x-\delta}{x})$ mutations to pop A.
(ii) Moving from x to $x + \delta$ takes $N(1 - \frac{V^*-x-\delta}{V^*-x})$ mutations to pop B.
(iii) Moving from x to δ takes $N(\frac{V^*-x}{V^*-\delta})$ mutations to pop B.
(iv) Moving from x to $V^* - \delta$ takes $N(\frac{x}{V^*-\delta})$ mutations to pop A.

Proof: This again follows immediately from Young's Lemma 1. All that needs be done is to note that both populations' 'sample size' is just N. □

Lemma 9

(i) If $\delta < x < V^* - \delta$, then moving from x to $x - \delta$ takes fewer mutations than moving from x to δ, and moving from x to $x + \delta$ takes fewer mutations than moving from x to $V^* - \delta$.

(ii) If $x = \delta$ then moving from x to 2δ takes the same number of mutations as moving from x to $V^* - \delta$.

(iii) If $x = V^* - \delta$ then moving from x to $V^* - 2\delta$ takes the same number of mutations as moving from x to δ.

Proof: Note that

$$\frac{V^* - x}{V^* - \delta} = 1 - \frac{x - \delta}{V^* - \delta} > (=) 1 - \frac{x - \delta}{x}, \tag{3}$$

as $V^* - \delta > (=) x$. The other half is shown by replacing x with y. □

A more exact statement of Proposition 4 is:

Lemma 10 *Let surplus be divided by the Nash demand game.*

(i) If $x^{NBS} < x^M$*, then the unique stochastically stable outcome is* $(I^*, V(I^*) - x^{NBS}, x^{NBS})$.

(ii) If $x^{NBS} \geq x^M$ *and* $\hat{r}(x^M) > r(x^M - \delta)$*, then the unique stochastically stable outcome is* $(I^*, V(I^*) - x^M, x^M)$.

(iii) If $x^{NBS} \geq x^M$ *and* $\hat{r}(x^M) < r(x^M - \delta)$*, then the unique stochastically stable outcome is* $(I^*, V(I^*) - x^M + \delta, x^M - \delta)$.

(iv) If $x^{NBS} \geq x^M$ *and* $\hat{r}(x^M) = r(x^M - \delta)$*, then the stochastically stable outcomes are* $(I^*, V(I^*) - x^M, x^M)$ *and* $(I^*, V(I^*) - x^M + \delta, x^M - \delta)$.

Proof: First assume that $x^L \leq x^{NBS}$. Construct Γ^* as follows: For each $x \in D(I^*)$ with $x < x^L$, find $\theta_x, \theta'_{x+\delta}$ such that $\rho(\theta_x) = \{(I^*, V^* - x, x)\}$, $\rho(\theta_{x+\delta}) = \{(I^*, V^* - x - \delta, x + \delta)\}$ and such that the cost of $(\theta_x, \theta_{x+\delta})$ is $r(x)$ (this is possible by Lemma 8.) Now find $\theta_{x^L}, \hat{\theta}$ such that $\rho(\theta_{x^L}) = \{(I^*, V^* - x^L, x^L)\}$, $\rho(\theta) = \{(\hat{I}, \hat{V} - \delta, \delta)\}$ and such that the cost of $(\theta_{x^L}, \hat{\theta})$ is $\hat{r}(x^L)$ (this is possible by Lemma 5.) Let Γ^0 include each of the above $(\theta_x, \theta_{x+\delta})$ (with $x < x^L$) as well as $(\theta_{x^L}, \hat{\theta})$. Let Θ^1 include each θ_x ($x \leq x^L$.) If $\hat{r}(x^L) > r(x^L - \delta)$, then let $\Gamma^1 = \Gamma^0 \setminus \{(\theta_{x^L}, \hat{\theta})\}$, and let $\theta^* = \theta_{x^L}$. If $\hat{r}(x^L) < r(x^L - \delta)$, then let $\Gamma^1 = \Gamma^0 \setminus \{(\theta_{x^L - \delta}, \theta_x^L)\}$, and let $\theta^* = \theta_{x^L - \delta}$. (In the case of equality, either construction will work.) Since Θ^1 contains an element from every locally stable set, if $\Theta^1 \subseteq \Theta^i \subset \bar{\Theta}$, then there exists $\theta_{i+1} \in \bar{\Theta} \setminus \Theta^i$ and $\theta' \in \Theta^i$ such that $r(\theta_{i+1}, \theta') = 1$. Now let $\Theta^{i+1} = \Theta^i \cup \{\theta_{i+1}\}$, and let $\Gamma^{i+1} = \Gamma^i \cup \{(\theta_{i+1}, \theta')\}$. Since there is a finite number of equilibria, there is some $\bar{i}$ such that $\Theta^{\bar{i}} = \bar{\Theta}$, at which point $\Gamma^* = \Gamma^{\bar{i}}$ is a θ^*-tree. Note that for any θ-tree Γ, if $\rho(\theta_1) \neq \rho(\theta)$ then $\exists \theta'_1, \theta_2$ such that $\rho(\theta_1) = \rho(\theta'_1) \neq \rho(\theta_2)$ and such that $(\theta'_1, \theta_2) \in \Gamma$. Now Γ^* has the least costly transition of this kind for each outcome other than $\rho(\theta^*)$, and since $r(x) < r(x + \delta)$, we know that any transition out of $\rho(\theta^*)$ is strictly more costly than transition out of any other outcome. Thus, θ is stochastically stable if and only if $\rho(\theta) = \rho(\theta^*)$ (unless $r(x^L - \delta) = \hat{r}(x^L)$ in which case either construction yields a lowest cost tree.)

The proof when $x^L > x^{NBS}$ proceeds exactly as above, except that in the construction of Θ^1 and Γ^1, for $x < x^{NBS}$ one needs $(\theta_x, \theta'_{x+\delta})$ and for $x^{NBS} < x \leq x^L$ one needs $(\theta_x, \theta'_{x-\delta})$. Note that $r(x) < r(x + \delta)$ if $x < x^{NBS}$ and $r(x) > r(x + \delta)$ if $x > x^{NBS}$. □

Finally, let us consider ultimatum bargaining. Some of the above lemmas for the Nash demand game carry through with minor modifications: Lemma 1 applies, except that there can clearly only be one set of demands following any investment. Obviously in an equilibrium, agents in population A accept all demands made in that equilibrium, so that if two agents in B were making different demands, then they would have different payoffs following that investment level, and one of them would imitate the other following updating.

Lemma 2 applies unchanged. Lemma 3 applies, but with the same caveat as in Lemma 1. Lemma 4 applies, but the proof must be changed so that agents in population B drift to believe that following an investment of I agents in population A will reject any demand greater than y. Next, we need some additional results.

Lemma 11 *Let surplus be divided by the ultimatum game. The component with the subgame perfect outcome, $(I^H, V^H - x^{\max}(I^H), x^{\max}(I^H))$, is a subset of the unique locally stable set.*

Proof: This is established by showing $\exists \theta^H$ such that $\theta^H \in T(\theta)$ for all equilibria θ and such that $\rho(\theta^H) = \{(I^H, V^H - x^{\max}(I^H), x^{\max}(I^H))\}$. To do this, consider an equilibrium θ which, by Lemma 3, we may assume to have single outcome, $\rho(\theta) = \{(I, V(I) - x, x)\}$. Let population A drift to expect a demand of $x^{\max}(I')$ following any investment $I' \neq I$, and to accept a demand of $x^{\max}(I)$ following I. Let a single agent in population B mutate to demand $x^{\max}(I)$. His demand will be accepted, and so if the rest of B update, then they will imitate him. At this point agents in A expect a maximal demand to follow all investments which makes (I^H, accept) their best choice. Hence when they update they will shift to this strategy, and we arrive at the desired outcome. $\square$

Lemma 12 *Let surplus be divided by the ultimatum game. Agents in population A receive a payoff of at least $V^H - I^H - x^{\max}(I^H)$ in every equilibrium.*

Proof: This is the worst payoff that an agent could expect for investing I^H. Hence if an agent in A were receiving less than this, he would change his investment to I^H. $\square$

Lemma 13 *Let surplus be divided by the 'ultimatum' game. If $V(I) - I - x \geq V^H - I^H - x^{\max}(I^H)$, then there exists an equilibrium θ such that $\theta \in \Theta^L$ and $\rho(\theta) = (I, V(I) - x, x)$.*

Proof: Immediate from Lemmas 11 and 4. $\square$

Note that of course if two outcomes give the same payoff to A, higher than that given by the hold–up equilibrium, then there are equilibria in which both outcomes are present. The above lemma is more a statement about the richness out equilibria, not a restriction. The latter is the job of the previous lemma.

Proof of Proposition 5: From Lemmas 11, 12 and 13 we know that this is the unique locally stable set, which Samuelson (1994) has shown must equal the stochastically stable set. $\square$

6 References

Binmore, Ken, Michele Piccione, and Larry Samuelson (1998): Evolutionary Stability in Alternating–Offers Bargaining Games, *Journal of Economic Theory* 80, 257–291.

Binmore, Ken, Chris Proulx, Larry Samuelson and Joe Swierzbinski (1998): Hard Bargains and Lost Opportunities, *Economic Journal* 108, 1279–1298.

Binmore, Ken, and Larry Samuelson (1999): Evolutionary Drift and Equilibrium Selection, *Review of Economic Studies* 66, 363–393.

van Damme, Erik, Reinhard Selten, and Eyal Winter (1990): Alternating Bid Bargaining with a Smallest Money Unit, *Games and Economic Behavior* 2, 188–201.

Ellingsen, Tore (1997): The Evolution of Bargaining Behavior, *Quarterly Journal of Economics* 112, 581–602.

Ellingsen, Tore and Magnus Johannesson (2000): Is There a Hold–up Problem? mimeo., Department of Economics, Stockholm School of Economics.

Fernandez, Raquel and Jacob Glazer (1991): Striking for a Bargain Between Two Completely Informed Agents, *American Economic Review* 81, 240–252.

Foster, Dean and H. Peyton Young (1990): Stochastic Evolutionary Game Dynamics, *Theoretical Population Biology* 38,219–232.

Fudenberg, Drew and David K. Levine (1993): Self–Confirming Equilibrium, *Econometrica* 61, 523–546.

Gale, John, Ken Binmore, and Larry Samuelson (1995): Learning to be Imperfect: The Ultimatum Game, *Games and Economic Behavior* 8, 56–90.

Grossman, Sanford J. and Oliver D. Hart (1986): The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration, *Journal of Political Economy* 94, 691–719.

Grout, Paul (1984): Investment and Wages in the Absence of a Binding Contract, *Econometrica* 52, 449–460.

Hackett, Steven C. (1994): Is Relational Exchange Possible in the Absence of Reputations and Repeated Contact? *Journal of Law, Economics and Organization* 10, 360–389.

Haller, Hans and Steinar Holden (1990): A Letter to the Editor on Wage Bargaining, *Journal of Economic Theory* 52, 232–236.

Hart, Oliver D. and John Moore (1999): Foundations of Incomplete Contracts, *Review of Economic Studies* 66, 115–138.

Holmström, Bengt and John Roberts (1999): The Boundaries of the Firm Revisited, *Journal of economic Perspectives* 12 (4), 73–94.

Kandori, Michihiro, George Mailath and Rafael Rob (1993): Learning, Mutation, and Long Run equilibria in Games, *Econometrica* 61, 29–56.

Klein, Benjamin, Robert G. Crawford, and Armen A. Alchian (1978): Vertical Integration, Appropriable Rents, and the Competitive Contracting Process, *Journal of Law and Economics* 21, 297–326.

Maskin, Eric and Jean Tirole (1999): Unforeseen Contingencies and Incomplete Contracts, *Review of Economic Studies* 66, 83–114.

Nash, John (1953): Two–Person Cooperative Games, *Econometrica* 21, 128–140.

Nöldeke, Georg and Larry Samuelson (1993): An Evolutionary Analysis of Backward and Forward Induction, *Games and Economic Behavior* 5, 425–454.

Rubinstein, Ariel (1982): Perfect Equilibrium in a Bargaining Model, *Econometrica* 50, 97–109.

Sáez–Martí, Maria and Jörgen Weibull (1999): Clever Agents in Young's Evolutionary Bargaining Model, *Journal of Economic Theory* 86, 268–279.

Samuelson, Larry (1994): Stochastic Stability in Games with Alternative Best Replies, *Journal of Economic Theory* 64, 35–64.

Ståhl, Ingolf (1972): *Bargaining Theory,* Stockholm: Economic Research Institute, Stockholm School of Economics.

Skyrms, Brian (1996): *Evolution of the Social Contract,* New York: Cambridge University Press.

TIROLE, JEAN (1986): Procurement and Renegotiation, *Journal of Political Economy* 94, 25–259.

TRÖGER, THOMAS (2000): Why Sunk Costs Matter for Bargaining Outcomes: An Evolutionary Approach, manuscript, ELSE, University College London.

WILLIAMSON, OLIVER (1975): *Markets and Hierarchies: Analysis and Antitrust Implications,* New York: The Free Press.

YOUNG, H. PEYTON (1993a): The Evolution of Conventions, *Econometrica* 61, 57–84.

YOUNG, H. PEYTON (1993b): An Evolutionary Model of Bargaining, *Journal of Economic Theory* 59, 145–168.

YOUNG, H. PEYTON (1998): Conventional Contracts, *Review of Economic Studies* 65, 773–792.