Mechanism Design with Renegotiation and Costly Messages

Robert Evans

March 2006

CWPE 0626

Not to be quoted without permission

Mechanism Design with Renegotiation and Costly Messages

Robert Evans

March 2006

Abstract

According to standard theory, the set of implementable outcome functions is reduced if the mechanism or contract can be renegotiated *ex post*. In some cases contracts can achieve nothing and so, for example, the holdup problem may be severe. This paper shows that if the mechanism is designed in such a way that sending a message involves a small cost (e.g., the opportunity cost of time spent attending a hearing) then *ex post* renegotiation essentially does not restrict the set of implementable functions. Any Pareto-efficient, bounded social choice function can be implemented in subgame-perfect equilibrium, for any strictly positive message cost.

1 Introduction

In one interpretation of the mechanism design problem, the agents involved are designing a mechanism for themselves, rather than having it imposed on them by an outside planner - the mechanism may, for example, be a contract or a constitution. In that case they will presumably choose a mechanism which will deliver a Pareto-efficient outcome in equilibrium. Suppose, however, that play of the mechanism results in an out-of-equilibrium outcome and that this outcome is Pareto-inefficient. Then, unless they have been able somehow to commit to this mechanism, the agents can be expected to tear up the contract and negotiate a new, Pareto-efficient outcome. In the two-agent case, this possibility of renegotiation limits the set of social choice rules which can be implemented, as shown by Maskin and Moore (1999) (see also Segal and Whinston (2002) and the survey in Maskin and Sjöström (2002)). The problem is that it may be necessary to punish an agent for deviating from the equilibrium, and moreover to do so without simultaneously rewarding the other agent. In that case, the mechanism would have to have inefficient outcomes for some combinations of messages, which is impossible if renegotiation cannot be prevented. The Maskin-Moore renegotiation paradigm has been influential in contract theory. For example, a number of well-known papers (e.g., Che and Hausch (1999), Segal (1999)) have shown in the context of buyer-seller models that, when *ex post* contract renegotiation cannot be prevented, there are plausible environments in which there is no contract which can improve on the null contract. These results in turn, it is argued, provide a foundation for the incomplete contracts paradigm.

One possible solution (Maskin-Moore (1999)) to the problem caused by *ex post* renegotiation is to include a third party. The two main parties can then both be punished by imposition of a fine which is paid to the third party, who need not be informed about the state. Because this outcome is efficient from the point of all three agents, it would not be renegotiated. One argument against this kind of scheme (see Hart and Moore (1999)) is that it would be vulnerable to collusion: for example one of the original two parties might collude with the third party and trigger the fine. On

the other hand, Baliga and Sjöström (2005) have shown that if collusive agreements and the original three-party agreement are modelled in a symmetric manner, and if the original mechanism may specify secret messages and secret cash transfers, then, in buyer-seller models and in moral-hazard-in-teams models, the first-best can be implemented in a collusion-free manner.

A second setting in which the possibility of renegotiation may not much constrain the set of implementable social choice rules is the case in which the two parties are risk-averse, as shown by Maskin and Moore (1999) (see also Maskin and Tirole (1999)). Suppose that it is possible for the mechanism to specify random outcomes after certain messages and that it is not possible for the parties to renegotiate in between the time that the messages are sent and the random variable is realized. Then, even though, for each possible realization of the lottery, renegotiation will take place to the efficient frontier, the *ex ante* payoffs can be inefficient because of the nonlinearity of the frontier. It is a matter of debate how practicable this scheme is - for example, whether it is possible to prevent renegotiation before the randomization.

In this paper, we show that even if the above solutions are not available, for whatever reason, the possibility of *ex post* renegotiation will still not constrain the implementable set if the mechanism can be designed in such a way that there is a strictly positive cost of sending a message (this cost can be arbitrarily close to zero). In the standard mechanism design model the parties can costlessly send messages (e.g., report their type) to the outside enforcer. In practice, however, sending any message will incur a strictly positive cost, if only the cost of making a telephone call or the opportunity cost of the time taken to attend a hearing. We show that if this friction is introduced to the model, then *ex post* renegotiation ceases to be a problem. More precisely, we consider a model with two risk-neutral agents, quasi-linear utility and complete information (i.e., the model encompasses the standard models found in the incomplete contracts literature). The result is that, for any strictly positive message cost, any Pareto-efficient, bounded social choice function which satisfies a weak preference-reversal condition can be strongly implemented in subgame-perfect equilibrium despite renegotiation.

To illustrate the argument, consider a simple buyer-seller model. There are two agents, a buyer (B) and a seller (S), and a single indivisible good. A *state of the world* is a pair (v, c), where v is B's utility for the good and c is S's cost of producing it. v and c are commonly known to B and S but not observable by the outside enforcer, i.e., not verifiable. A Pareto-efficient, bounded social choice function f specifies that the good should be produced and traded if and only if $v \geq c$ and specifies, as a function of the state (v, c), a bounded money transfer (possibly negative) to be paid by B to S. If the parties make investments before production which affect the state then, to achieve efficiency (avoid the hold-up problem), f will in general have to make the transfer sensitive to the state; the problem is that renegotiation may make this impossible. Ideally, we would like to know that every such f can be implemented. Assume that the agents split equally any gains from renegotiation. To implement f we can use a mechanism with the following features.

First S announces the state, i.e., a pair (v_s, c_s). B can then either say nothing or else make a challenge. If he says nothing, the output of the mechanism is the outcome $f(v_s, c_s)$. A challenge by B consists of a nomination of one of four possible dispute procedures, corresponding respectively to a claim by B that $v < v_s, v > v_s, c < c_s$ and $c > c_s$. For illustration, consider the first of these procedures. B names a price $p < \frac{v_s}{2} - \epsilon$, where $\epsilon > 0$ is small compared to the message cost. S pays a large sum K to B. S then, at each of a large number T of stages, chooses either to send no message, thereby triggering the default outcome 'produce and trade the good for price p', or else to send a message requesting a delay, in which case she chooses again at the next stage. If S requests a delay T times then the outcome is 'produce the good but do not transfer it, and make no money transfer'. Each time that S delays, B has to pay S the message cost, less ϵ.

Consider the case $v > c$, in which trade is efficient. Suppose that stage T has been reached. S's payoff if she does not delay is $p - c$, while if she does delay she gets $\frac{v}{2} - c - \epsilon$ since her delay cost[1] is ϵ and we assume that the renegotiation surplus of v will be equally split. Thus, if S has announced $v_s > v$ and B has nominated

---

[1] We assume that there is no discounting.

price p such that $\frac{v_s}{2} - \epsilon > p > \frac{v}{2} - \epsilon$, S will choose not to delay at stage T and, by backward induction, will choose not to delay at every previous stage, if reached. The transfer of K gives B the incentive to challenge, and S is therefore deterred from misrepresenting the state.

The question is, why doesn't B make an invalid challenge, e.g. claim that $v < v_s$ when in fact $v = v_s$? If he does so, then, at stage T, S will choose to delay since $p < \frac{v}{2} - \epsilon$. By backward induction she will also choose to delay at every earlier stage. We assume that the parties can renegotiate during the play of the mechanism as well as at the end, so that they will in fact reach agreement at the outset in order to avoid the wasteful message costs. Nevertheless, since B is obliged to bear most of these costs his bargaining position is weak and so, if T is large enough, he is heavily punished for an invalid challenge[2].

The reason that the framework of this paper delivers different results from those derived in the Maskin-Moore framework is that in the latter every possible play of the mechanism is *ex ante* efficient. Here, by contrast, once a message has been sent, some resources have been used up. Subsequent renegotiation ensures that the *equilibrium continuation play* will be efficient but it cannot recover the sunk cost, so such an outcome will not be efficient *ex ante*.

The result described here is related to the work of Watson (2001, 2004), which demonstrates that mechanism design theory needs to take account of the interaction between renegotiation opportunities and the technology of trade. Watson distinguishes between *public-action* models and *individual-action* models. A public action is one taken by the external enforcement authority, while an individual action is an inalienable decision of one of the contracting parties. The standard theory of mechanism design treats actions as public (alternatively, actions are individual but the analysis is limited to *forcing contracts*). Watson shows that in an *interim* renegotiation setting (i.e., one in which renegotiation takes place only before the messages are sent) public-action and individual-action models are equivalent. However, in an *ex post* renegotiation setting more functions can be implemented in an individual-action

---

[2]Note that there is no point in S threatening to delay if the challenge is valid because both players know that she will strictly prefer not to delay when she has to make the decision.

model. Therefore, given the possibility of *ex post* renegotiation, explicit modeling of the technology of trade is necessary. Sending a message is necessarily an individual action; if doing so is costly then, in an *ex post* renegotiation setting, one would expect that explicit modeling of these costs is necessary. The result derived in this paper is that, if these costs are modeled, then the interim and *ex post* renegotiation settings turn out to be equivalent in the standard complete information model.

Another paper which is close to the present one is Rubinstein and Wolinsky (1992), which shows that if a time dimension is added to an implementation problem and renegotiation is modeled as costly because it involves delay then the set of implementable outcomes is expanded, compared with those of the standard model of implementation with renegotiation. There are a number of differences from the current paper. Firstly, Rubinstein and Wolinsky establish results only for a version of the buyer-seller model with one good in which trade is always efficient. Secondly, the only implementable price functions $p(v, c)$ are those which are monotonic and satisfy $v > p(v, c) > c$. The results in this paper are therefore much stronger. Furthermore, the approach is somewhat different. The mechanism that Rubinstein and Wolinsky use is a bargaining game with discounting and the renegotiation-proofness criterion is that at no stage should the players both be able to benefit by substituting a different outcome one period in the future. This seems to require significant ability of the outside enforcer to structure the negotiations.[3] The approach of the current paper is that the bargaining game is exogenous and the mechanism has to be designed taking this given game into account. Furthermore, renegotiation is costless.

The dominant model of mechanism design is elegant and tractable, but it is not the only possible model, and in certain circumstances it leaves out some important considerations. Clearly, it is a common feature of many real contracts that they include provision for dispute resolution processes which are time-consuming and, therefore, necessarily costly. The parties to such a contract will have preferences over the manner in which the mechanism plays out as well as over the final outcome and this makes

---

[3]For example, as part of the mechanism, it is possible for one player to make a proposal in a certain set, the other to accept it, and the payoffs to be realized, all within the same period, but it is not possible for them to agree within that period on a different (renegotiation) proposal.

a difference to the analysis, particularly, as we show, when renegotiation cannot be prevented.

There are relatively few papers which study implementation theory with preferences over process as well as outcomes. Rubinstein and Wolinsky (1992) has been discussed above. Another that does so is Glazer and Rubinstein (1998), in which voters care about how many others vote with them as well as about the outcome. Also relevant is a group of papers which examine models of mechanism design in which agents have different costs of sending different kinds of messages, for example because evidence has to be produced, or because there is a psychic cost of lying. Examples are Green and Laffont (1986), Bull and Watson (2004), and Deneckere and Severinov (2001). The message costs in the current paper are different because they do not depend on the type of the sender.

2 The Model

There are two agents, 1 and 2, both risk-neutral. The environment is $\langle D, \Theta \rangle$, where D is the set of *outcomes* and Θ is the set of *states of the world*. A feasible outcome consists of a physical outcome (or *action*) in a set A together with a pair of money transfers (y_1, y_2) with non-positive sum. That is,

$$D = \{(a, y_1, y_2) \in A \times \Re^2 | y_1 + y_2 \leq 0\}.$$

We discuss below the case in which the set of actions depends on the state, but for the moment A is assumed to be the same in all states. Each agent has a payoff function which depends on the state and is quasi-linear in money: if the action is $a \in A$, the state is $\theta \in \Theta$, and agent i's ($i \in \{1, 2\}$) money payment is $y_i \in \Re$ then i's payoff is $u_i(a, \theta) + y_i$. We assume that the u_i functions are bounded: there exists a number M_1 such that, for $i \in \{1, 2\}$ and all $(a, \theta) \in A \times \Theta$, $|u_i(a, \theta)| < M_1$. The assumptions of risk-neutrality and quasi-linearity are made partly for expositional reasons and partly because these are the standard assumptions in the incomplete contracts literature. The result does not depend on them.

A *social choice function* (SCF) is a function $f : \Theta \to D$. $f(\theta) = (a^f(\theta), y_1^f(\theta), y_2^f(\theta))$ is the socially-optimal (or f-optimal) outcome when the state is θ. For simplicity, we assume that in each state $\theta \in \Theta$ there is a unique Pareto-efficient action $a^*(\theta)$. That is, $a^*(\theta)$ is the unique solution to the problem

$$max_{a \in A} u_1(a, \theta) + u_2(a, \theta).$$

The maximized value of this problem is denoted by $\sigma(\theta)$.

A social choice function f is *Pareto-efficient* if and only if for each $\theta \in \Theta$, $a^f(\theta) = a^*(\theta)$ and $y_1^f(\theta) + y_2^f(\theta) = 0$. It is *bounded* if and only if there exists M_2 such that, for $i \in \{1, 2\}$ and all $\theta \in \Theta$, $|y_i^f(\theta)| < M_2$.

Once the state of the world is realized, this true state is common knowledge among the agents (i.e., there is *complete information*). The agents then play a mechanism. We consider extensive-form mechanisms: thus, a mechanism is a finite game-form in extensive form. Since we are interested in implementation with renegotiation, we will consider mechanisms which consist of an *underlying mechanism* together with renegotiation moves. At each decision node in the underlying mechanism, an agent chooses a message from some specified set of messages, or else sends no message, and each terminal node corresponds to an outcome in D, this outcome being common knowledge. If, given the true state θ, this outcome is inefficient, then we assume that before it is enforced the players have time to renegotiate to an efficient outcome. We make the assumption, common in the incomplete contracts literature, that the surplus from renegotiation is split equally, so if the outcome of the mechanism is (a, y_1, y_2) and the state is θ then the players agree to enforce the outcome $(a^*(\theta), y, -y)$ where y is chosen so that player i's utility for this outcome is

$$u_i(a, \theta) + y_i + \frac{1}{2}[\sigma(\theta) - u_1(a, \theta) - u_2(a, \theta) - y_1 - y_2].$$

That is,

$$y = \frac{1}{2}[(u_2(a^*(\theta), \theta) - u_2(a, \theta)) - (u_1(a^*(\theta), \theta) - u_1(a, \theta)) + (y_1 - y_2)].$$

8

To this point the framework is the same as that of Maskin and Moore (1999), for the case of quasi-linear payoff functions and the specific renegotiation rule just given. We now introduce two modifications to their framework. The first, which is minor, is that we assume that renegotiation can take place not only at the end of the mechanism, but also during the play of the mechanism. Since the mechanism may have many stages it is natural to suppose that if the players at any stage expect the continuation play to be inefficient then they will immediately renegotiate to an efficient outcome. To model this, we assume that immediately after every decision node of the underlying mechanism there is a renegotiation stage which always takes the following form. First, player 1 proposes an outcome $d_1 \in D$ to player 2. Player 2 then either accepts, in which case d_1 is the final outcome and the remainder of the mechanism is abandoned; or 2 rejects, in which case one of the two players is chosen at random, with equal probabilities, to make a second proposal $d_2 \in D$. If the non-proposer accepts then d_2 is the final outcome; if not, this renegotiation phase is over. If a renegotiation phase ends without acceptance of a proposal then the players move to the next message stage of the mechanism, or, if the final message stage has passed, the outcome stipulated by the mechanism is implemented. It is easy to see that a subgame-perfect equilibrium (SPE) of this bargaining game, when played after a final node of the underlying mechanism, will give the equal-sharing renegotiation rule set out in the previous paragraph. Similarly, if, at some message stage, there is a unique SPE continuation, and this is inefficient, then renegotiation will take place at the preceding bargaining stage, with the continuation payoffs acting as disagreement payoffs. We also assume that there is a renegotiation phase immediately before the mechanism is played (after the players learn θ)[4].

The second modification is that we assume that it costs a strictly positive amount k to send a message[5], where k may be arbitrarily small[6]. This cost enters the sender's

---

[4]In the standard model the initial renegotiation would be redundant, given the possibility of *ex post* renegotiation. We shall see that this is not so in the present model because messages are costly.

[5]It is not necessary to assume that sending a message is *necessarily* costly, only that the mechanism can be designed in such a way that certain messages have to be sent in a way which imposes a cost.

[6]I owe to Joel Watson the suggestion that small costs may be sufficient.

payoff function in the same way as a deduction of k units of money. We assume that the players do not discount the future, so it makes no difference whether the cost is deducted at the time the message is sent or at the end of the game. Our convention is that it is deducted at the time at which the message is sent: thus, when we refer below to a 'continuation payoff', this is to be understood as excluding all message costs (and transfer payments) incurred in the history to date. There is assumed to be no cost of renegotiating the mechanism. Since it is the possibility of renegotiation which limits what can be achieved, it strengthens the result of this paper if renegotiation is assumed to be as easy as possible.

We refer to a mechanism with renegotiation as described above and with message cost k as a *k-mechanism*.

Definition   A social choice function f is *k-implementable in subgame-perfect equilibrium with renegotiation* if there exists a k-mechanism such that, for each $\theta \in \Theta$, there exists a subgame-perfect equilibrium of the mechanism and the outcome of every SPE is $f(\theta)$.

Player 1 prefers $(a, y, -y)$ to $(a', y', -y')$ in state θ if

$$u_1(a, \theta) + y + \frac{1}{2}[\sigma(\theta) - u_1(a, \theta) - u_2(a, \theta)]$$

$$> u_1(a', \theta) + y' + \frac{1}{2}[\sigma(\theta) - u_1(a', \theta) - u_2(a', \theta)],$$

i.e., if

$$[u_1(a, \theta) - u_1(a', \theta)] - [u_2(a, \theta) - u_2(a', \theta)] > 2(y - y').$$

It follows that it is possible to find (y, y') such that 1 prefers $(a, y, -y)$ to $(a', y', -y')$ in state θ and vice versa in state θ', i.e., preference reversal holds allowing for renegotiation, only if $d(a, a', \theta) > d(a, a', \theta')$ where

$$d(a, a', \theta) = [u_1(a, \theta) - u_1(a', \theta)] - [u_2(a, \theta) - u_2(a', \theta)].$$

Therefore we will need to make the following assumption.

10

Assumption 1: Given $\theta, \theta' \in \Theta$ such that $f(\theta) \neq f(\theta')$, there exist $a \in A$ and $a' \in A$ such that

$$d(a, a', \theta) > d(a, a', \theta'). \tag{1}$$

Assumption 1 is fairly weak and indeed is automatically satisfied in the buyer-seller models found in the incomplete contracts literature. To see this, consider the following model.

Example: Buyer-Seller Model Player 1 is a seller (S) and player 2 is a buyer (B), who will produce and trade at most one unit of some indivisible good. The set of goods which may potentially be produced is G; S will produce at most one of these. The set of states of the world $\Theta = V \times C$ where V and C are sets of real-valued functions of G. In state (v, c), $v(g)$ is B's value for good $g \in G$ and $c(g)$ is S's cost of producing g. The set of actions $A = \{G \times \{1, 0\}\} \cup \emptyset$, where $(g, 1)$ means that S produces $g \in G$ and transfers it to B, $(g, 0)$ means that S produces g but does not transfer it to B, and $\emptyset$ means that no good is produced. If $\theta = (v, c)$ then $u_B((g, 1), \theta) = v(g)$, $u_B((g, 0), \theta) = u_B(\emptyset, \theta) = 0$, $u_S((g, 1), \theta) = u_S((g, 0), \theta) = -c(g)$, and $u_S(\emptyset, \theta) = 0$.

If $(v, c) = \theta \neq \theta' = (v', c')$ then either $v(g) \neq v'(g)$ for some $g \in G$ or $c(g) \neq c'(g)$ for some $g \in G$, or both. Suppose that $v(g) > v'(g)$. If we take $a = (g, 0)$ and $a' = (g, 1)$ then $d(a, a', \theta) = v(g)$ and $d(a, a', \theta') = v'(g)$, so (1) is satisfied. If $v(g) < v'(g)$, take $a = (g, 1)$ and $a' = (g, 0)$, giving

$$-v(g) = d(a, a', \theta) > d(a, a', \theta') = -v'(g).$$

If $v(g) = v'(g)$ but $c(g) \neq c'(g)$ then we can take $(a, a') = (\emptyset, (g, 0))$ if $c(g) > c'(g)$ and $(a, a') = ((g, 0), \emptyset)$ if $c(g) < c'(g)$. Again, (1) is satisfied in each case.

3 The Implementation Result

In the framework set out above, but without message costs, the set of efficient

implementable outcome functions would be severely restricted, as shown, for example, by Maskin and Moore (1999), Segal (1999) and Che and Hausch (1999). But, as the following result shows, matters are different if message costs are positive.

THEOREM: *Any bounded, Pareto-efficient social choice function which satisfies Assumption 1 is k-implementable in subgame-perfect equilibrium with renegotiation for any $k > 0$.*

We provide here an outline of the proof, the remaining details being in the Appendix. Take an SCF f which satisfies the conditions of the Theorem. Let M_2 be an upper bound on $|y_i^f(\theta)|$. For each ordered pair (θ, θ') let $a(\theta, \theta')$ and $a'(\theta, \theta')$ be actions satisfying (1) in Assumption 1. Let K be a number satisfying

$$K > 4M_1 + M_2 + k, \tag{2}$$

let ϵ satisfy $0 < \epsilon < \frac{k}{2}$, let T be an integer satisfying

$$T > \frac{4K}{k - 2\epsilon}, \tag{3}$$

and, finally, let

$$y(\theta, \theta') = \frac{1}{4}[d(a(\theta, \theta'), a'(\theta, \theta'), \theta) + d(a(\theta, \theta'), a'(\theta, \theta'), \theta'] - \epsilon. \tag{4}$$

(Where the pair (θ, θ') under consideration is clear, we will drop the arguments and refer to a, a' and y.)

The following describes the underlying mechanism which will implement f (underlying because the renegotiation stages are omitted in the description).

1. Player 1 announces a state $\theta \in \Theta$.

2. Player 2 either (a) sends no message, in which case the mechanism ends with $(a^*(\theta), y_1^f(\theta) + \frac{k}{2}, y_2^f(\theta) - \frac{k}{2})$ as the outcome; or (b) challenges and names a state θ' such that $f(\theta') \neq f(\theta)$. If 2 challenges then 1 pays K to 2 and there begins a challenge procedure of up to T stages, as follows.

In stage t of the challenge procedure ($t = 1, .., T$) player 1 either (i) sends no message, in which case the mechanism ends with outcome $(a', 0, 0)$; or (ii) sends the message 'wait', in which case 2 pays $k - \epsilon$ to 1 and play moves to stage $t + 1$, unless $t = T$ in which case the mechanism ends with outcome $(a, -y, y)$.

If a player deviates from the above rules, he or she pays a large fine to the other player.

This completes the description of the mechanism. We refer to this mechanism, including the renegotiation stages, as $M^f(K, T, \epsilon)$. We show in the Appendix that in any SPE of $M^f(K, T, \epsilon)$:

(a) if the true state is θ, then in the subgame in which 1 has announced θ and 2 has just challenged with $\theta' \neq \theta$, player 2's continuation payoff is

$$K + \frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a, \theta) - u_2(a, \theta)] + y + T\epsilon - \frac{Tk}{2}; \tag{5}$$

(b) if the true state is θ', then in the subgame in which 1 has announced $\theta \neq \theta'$ and 2 has just challenged with θ', player 2's continuation payoff is

$$K + \frac{\sigma(\theta')}{2} - \frac{1}{2}[u_1(a', \theta') - u_2(a', \theta')]. \tag{6}$$

The reasoning behind (a) is that if stage T is reached when the challenge was invalid, then 1 prefers to say 'wait', i.e., choose $(a, -y, y)$, by (1). Of course, renegotiation takes place before she does so (after the previous message) in order to avoid the loss k associated with the message. By backward induction, if no renegotiation has taken place by stage $t < T$ then 1 will say 'wait' there too. At the renegotiation which takes place immediately after the challenge, the cumulative gain, Tk, from not going down this path is shared equally between the players. Nevertheless, as expression (5) shows, 2 bears a substantial cost from an invalid challenge because the mechanism requires him to bear most of 1's message costs. The argument for (b) is more straightforward. If stage T of the challenge procedure is reached after a valid challenge, player 1 will prefer to choose $(a', 0, 0)$ (i.e., 'no message'). By backward induction, she strictly prefers to make the same choice at any earlier stage, rather

13

than incur cost ϵ and get $(a', 0, 0)$ at the next stage. So, if the challenge is valid (1 has lied), the mechanism ends at the first stage of the challenge procedure with 2 getting the payoff in expression (6).

Suppose θ is true and 1 has told the truth (announced θ). If 2 does not challenge 2 gets

$$u_2(a^f(\theta), \theta) + y_2^f(\theta),$$

while, by (5), if he challenges he gets

$$K + \frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a, \theta) - u_2(a, \theta)] + y + T\epsilon - \frac{Tk}{2} - k;$$

which, by (2) and (3), is strictly less. This shows that, once 1 has announced the true state, 2 cannot profitably challenge.

Now suppose that θ' is true but 1 has announced θ. If 2 does not challenge, he gets

$$u_2(a^f(\theta), \theta') + y_2^f(\theta),$$

while, by (6), if he challenges he gets

$$K + \frac{\sigma(\theta')}{2} - \frac{1}{2}[u_1(a', \theta') - u_2(a', \theta')] - k,$$

which is strictly greater by (2). This shows that 2 will always challenge a false announcement by 1.

Now consider 1's choice of announcement at the start. Suppose that the true state is θ'. If 1 announces θ' then her payoff is

$$u_1(a^f(\theta'), \theta') + y_1^f(\theta') - k \tag{7}$$

since 2 will not challenge. If she announces $\theta \neq \theta'$ then 2 will challenge. If he challenges with θ' then player 2's payoff will be given by (6). Since the total surplus

14

available is $\sigma(\theta')$, 1's payoff if she announces θ is therefore bounded above by

$$\frac{\sigma(\theta')}{2} - K + \frac{1}{2}[u_1(a', \theta') - u_2(a', \theta')], \tag{8}$$

which is less than the expression in (7) by (2). This shows that in any SPE player 1 tells the truth.

Thus, the unique SPE outcome in state θ, starting from the first announcement stage, is that 1 announces θ, incurring cost k, 2 does not challenge, there is no renegotiation and the final outcome, net of the message cost, is $(a^*(\theta), y_1^f(\theta) - \frac{k}{2}, y_2^f(\theta) - \frac{k}{2})$. Therefore, once θ is realized, the players renegotiate (before playing the mechanism) to split the surplus of k, giving outcome $(a^*(\theta), y_1^f(\theta), y_2^f(\theta)) = f(\theta)$. This proves the Theorem.

4   Discussion

Some continuation equilibria of the mechanism used here are inefficient. For example, at some nodes a player has to decide whether or not to send a message, and in some circumstances his equilibrium strategy is to do so. A strong version of renegotiation-proofness (one which is common in the literature) would rule this out, the assumption being that somehow the players would renegotiate out of such a continuation. However, the mechanism above specifies that there are specific dates at which a message must be sent in order to avoid a particular consequence. If that date is reached and no renegotiation has yet happened, then it is too late to renegotiate. The relevant player must take whatever action is optimal for himself at that point. Of course, the players anticipate this inefficient outcome and renegotiate beforehand in order to avoid it.

The assumption above is that the surplus is equally shared in renegotiation, but the analysis can easily be adapted to other specifications. If player 2 has all the bargaining power then the mechanism above will not work because it would not be costly for him to invoke the dispute procedure, but in that case one could use a mechanism in which the roles of the two players are interchanged. Similarly, the formal model of

15

bargaining outlined above assumes, since it is a finite-stage procedure, that if bargaining ends without agreement then there is an inefficient outcome. This is incompatible with the idea of renegotiation-proofness, but it is easy to see that the analysis can be adapted to include an infinite-horizon protocol. Assume that the parties discount the future and bargain according to an infinite-horizon alternating-offers model. If the periods are sufficiently close together then this model will approximate to the one analyzed above and the result will go through.

The mechanism $M^f(K, T, \epsilon)$ (like many similar mechanisms found in the implementation literature) could be criticized on the grounds that it asks a player to describe the state θ, which might be very costly. On the other hand, in equilibrium the state is not actually described because renegotiation takes place beforehand in order to save on the costs of doing so. Even if describing the entire state is infeasible, for many models it would suffice to have a mechanism in which a small subset of the information contained in the state is described. For example, for the buyer-seller models in the literature, one would only need to describe the name of the efficient action (good) and the utility pair associated with it. In that case the challenge procedure would have to involve the option of making a counter-challenge in which a different action is claimed to be efficient. For discussion of this issue, see Maskin and Tirole (1999).

Similarly, if the action space A depends on the state θ then the result will still obtain. In that case the mechanism would specify that after 1 has announced θ, 2 can, in addition to challenging the reported utilities, challenge the implied report about which actions are feasible. That is, 2 can exhibit an action which is infeasible though 1 has reported that it is feasible, or vice versa. By the usual assumptions, such statements are directly verifiable by the outside enforcer, and 2 can be rewarded for a valid challenge.

APPENDIX

This Appendix provides the remainder of the proof of the Theorem. Take an

SCF f which satisfies the conditions in the statement of the Theorem, and define the mechanism $M^f(K, T, \epsilon)$ as in the main text. It remains to show that in any SPE of $M^f(K, T, \epsilon)$:

(a) if the true state is θ, then in the subgame in which 1 has announced θ and 2 has just challenged with θ', player 2's continuation payoff is

$$K + \frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a, \theta) - u_2(a, \theta)] + y + T\epsilon - \frac{Tk}{2}; \tag{9}$$

(b) if the true state is θ', then in the subgame in which 1 has announced $\theta \neq \theta'$ and 2 has just challenged with θ', player 2's continuation payoff is

$$K + \frac{\sigma(\theta')}{2} - \frac{1}{2}[u_1(a', \theta') - u_2(a', \theta')]. \tag{10}$$

(a) Consider a subgame in which the state is θ, 1 has announced θ, i.e. told the truth, and 2 has challenged with state $\theta' \neq \theta$. Suppose that stage T of the challenge procedure has been reached and there has been no renegotiation so far. If 1 sends no message then the outcome of the mechanism will be $(a', 0, 0)$ which will be renegotiated so as to give 1 a payoff of

$$\frac{\sigma(\theta)}{2} + \frac{1}{2}[u_1(a', \theta) - u_2(a', \theta)]. \tag{11}$$

If instead she sends the message 'wait' then, after renegotiation, her payoff (net of message cost) will be

$$\frac{\sigma(\theta)}{2} + \frac{1}{2}[u_1(a, \theta) - u_2(a, \theta)] - y - \epsilon \tag{12}$$

so she prefers to say 'wait' if

$$d(a, a', \theta) > 2y + 2\epsilon. \tag{13}$$

From (4) and Assumption 1, this is satisfied, so 1 chooses to wait. Therefore there is a unique SPE continuation and 1's continuation payoff at (or just before) the stage-T

choice is made is given by (12) while player 2's is

$$\frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a,\theta) - u_2(a,\theta)] + y + \epsilon - k. \tag{14}$$

Now suppose, as an induction hypothesis, that just before the decision at stage t ($t = 2, ..., T$), when no renegotiation has taken place so far, there is a unique SPE continuation, 1's continuation payoff is

$$\frac{\sigma(\theta)}{2} + \frac{1}{2}[u_1(a,\theta) - u_2(a,\theta)] - y - (T - t + 1)\epsilon + (T - t)\frac{k}{2} \tag{15}$$

and 2's is

$$\frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a,\theta) - u_2(a,\theta)] + y + (T - t + 1)\epsilon - (T - t)\frac{k}{2} - k. \tag{16}$$

This is true for $t = T$ by (12) and (14). Expressions (15) and (16) sum to $\sigma(\theta) - k$ so there is a renegotiation surplus of k after the message 'wait' has been sent at stage $t - 1$, which will be equally split. Thus, at stage $t - 1$, 1 chooses between 'no message', giving the payoff in expression (11), and 'wait', which gives her

$$\frac{\sigma(\theta)}{2} + \frac{1}{2}[u_1(a,\theta) - u_2(a,\theta)] - y - (T - t + 2)\epsilon + (T - t)\frac{k}{2} + \frac{k}{2}, \tag{17}$$

since she has to bear cost ϵ if she waits. Thus, she prefers 'wait' if

$$d(a, a', \theta) > 2y + 2\epsilon - (k - 2\epsilon)(T - (t - 1))$$

which is true by (13); hence, in the unique SPE continuation, 1's payoff at stage $t - 1$ is given by (17) and 2's is

$$\frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a,\theta) - u_2(a,\theta)] + y + (T - t + 2)\epsilon - (T - (t - 1))(\frac{k}{2}) - k,$$

which proves the induction hypothesis. Therefore, (using (16) with $t = 1$) once 2 has challenged, his continuation payoff in the unique SPE continuation, including the

payment K and adding $\frac{k}{2}$ for the renegotiation which takes place immediately after the challenge, is

$$K + \frac{\sigma(\theta)}{2} - \frac{1}{2}[u_1(a,\theta) - u_2(a,\theta)] + y + T\epsilon - \frac{Tk}{2}.$$

(b) Consider a subgame in which 1 has announced θ, 2 has challenged with θ', but the true state is θ', so the challenge is valid.

By the logic leading to (13), at stage T of the challenge procedure 1 strictly prefers to send no message if

$$2y + 2\epsilon > d(a, a', \theta'),$$

which is true by (4) and Assumption 1. Suppose that at stage $t > 1$ player 1 strictly prefers to send no message. This continuation being efficient (allowing for the renegotiation from outcome $(a', 0, 0)$), there is no renegotiation after the message 'wait' at stage $t-1$, so 1 chooses at stage $t-1$ between (i) pre-renegotiation outcome $(a', 0, 0)$ now or (ii) the same outcome at the following stage, less cost ϵ. Clearly the former is better. Therefore, by backward induction, the unique equilibrium payoff for 2 if he challenges with θ' is

$$K + \frac{\sigma(\theta')}{2} - \frac{1}{2}[u_1(a', \theta') - u_2(a', \theta')],$$

which establishes (10). QED.

REFERENCES

Baliga, S. and T. Sjöström (2005), "Contracting with Third Parties", mimeo, MEDS.

Bull, J. and J. Watson (2004), "Evidence Disclosure and Verifiability", *Journal of Economic Theory*.

Che, Y-K., and D. Hausch (1999), "Cooperative Investments and the Value of Contracting", *American Economic Review*, 89, 125-147.

19

Deneckere, R. and S. Severinov (2003), "Mechanism Design and Communication Costs", mimeo, University of Wisconsin.

Glazer, J., and A. Rubinstein (1998), "Motives and Implementation: on the Design of Mechanisms to Elicit Opinions", *Journal of Economic Theory*, 79, 157-173.

Green, J. and J.-J. Laffont (1986), "Partially Verifiable Information and Mechanism Design", *Review of Economic Studies*, 53, 447-456.

Hart, O. and J. Moore (1999), "Foundations of Incomplete Contracts", *Review of Economic Studies*, 66, 115-138.

Maskin, E. and J. Moore (1999), "Implementation and Renegotiation", *Review of Economic Studies*, 66, 39-56.

Maskin, E. and J. Tirole (1999), "Unforeseen Contingencies and Incomplete Contracts", *Review of Economic Studies*, 66, 83-113.

Maskin, E. and T. Sjöström (2002), "Implementation Theory", in K. J. Arrow, A. K. Sen and K. Suzumura, (eds.), *Handbook of Social Choice and Welfare*, Elsevier.

Rubinstein, A. and A. Wolinsky (1992), "Renegotiation-Proof Implementation and Time Preferences", *American Economic Review*, 82, 600-614.

Segal, I. (1999), "Complexity and Renegotiation: A Foundation for Incomplete Contracts", *Review of Economic Studies*, 66, 57-82.

Segal, I. and M. Whinston (2002), "The Mirrlees Approach to Mechanism Design with Renegotiation", *Econometrica*, 70, 1-47.

Watson, J. (2001), "Contracts, Mechanism Design and Technological Detail", mimeo, UCSD.

Watson, J. (2004), "Contract and Game Theory: Basic Concepts for Settings with Finite Horizons", mimeo, UCSD.