Baliga, Sandeep; Sjöström, Tomas

Working Paper

Contracting with third parties

CSIO Working Paper, No. 0075

THE CENTER FOR THE STUDY
OF INDUSTRIAL ORGANIZATION
AT NORTHWESTERN UNIVERSITY

Working Paper #0075

Contracting with Third Parties*

By

Sandeep Baliga
Kellogg School of Management

&

Tomas Sjöström
Rutgers University

This Draft: October 9, 2005

* We thank Joel Watson and seminar audiences at Haas (Berkeley), NYU and Chicago GSB for helpful comments.

Contracting with Third Parties

Sandeep Baliga
Kellogg SM (MEDS)
2001 Sheridan Road
Evanston, IL 60208.

Tomas Sjöström
Department of Economics
Rutgers University
New Brunswick, NJ 08901.

This version: October 9, 2005*

Abstract. In the bilateral hold-up model and the moral hazard in teams model, introducing a third party allows implementation of the first-best outcome, even if the agents can renegotiate inefficient outcomes and collude. Fines paid to the third party provide incentives for truth-telling and first-best levels of investment. Our results suggest that models that provide foundations for hold-up and incomplete contracts by invoking renegotiation are sensitive to the introduction of third parties.

1. Introduction

This paper investigates the role of third parties in contracting problems. Canonical hold-up models typically assume that a buyer deals with a seller, but third parties are not allowed, or perhaps dismissed in a footnote (e.g., Hart and Moore [9]). In the implementation literature, the two-agent case is often treated prominently, even though it is known that implementation is easier when a third party is added (e.g., Maskin and Moore [14]). The third party can be useful even if he has no information about the state of the world, because he can be a "budget-breaker", as discussed by Holmström [12].

Suppose two parties, an architect and a builder, cooperate to build a building. The quality of the building will depend on three things: the quality of the architect's design, the builder's skill, and a stochastic shock. These three variables are the state of the world. The two parties learn the true state, but no outsider can observe it (although an outsider may be able to judge the quality of the building after it is built). Suppose the contract specifies ex post transfers as a function of some announcements made by the two parties (a "message game"). If both report the state truthfully, the transfers will reflect the contributions of each party and thereby provide correct incentives to invest in the transaction. However, in order to provide an incentive for them to tell the truth, it may be necessary to punish both of them if they disagree. In

*We thank Joel Watson and seminar audiences at Haas (Berkeley), NYU and Chicago GSB for helpful comments.

the absence of a third party, this punishment typically involves an ex post inefficiency (say, the destruction of the building). If such outcomes can be *renegotiated*, then it may be impossible to implement the first best. This logic underlies models of the hold-up problem which provide foundations for "incomplete contracts", e.g., Che and Hausch [3] and Segal [19]. But with a third party, ex post renegotiation becomes less of a problem, because the punishment can consist of fines paid to the third party. (Since the fines do not waste any resources, the outcome is ex post efficient.) This suggests that results derived in bilateral hold-up models, which rely on renegotiation to generate inefficiencies, may not be robust to the introduction of third parties.

It is often argued that contracts with a third party are vulnerable to collusion. For example, the third party might collude with the builder to extract a fine from the architect. However, in this paper we show that it is possible to design the original contract so that no collusion occurs in equilibrium. We consider two standard contracting models: the buyer-seller model, and the moral hazard in teams model. In these environments, contracts can be designed to eliminate collusion and implement the first-best outcome, even if the agents can renegotiate inefficient outcomes.

Following the pioneering work of Tirole [22], much of the literature has assumed that subcoalitions can use a more powerful contracting technology than the grand coalition (e.g., footnote 20 in Hart and Moore [9]). Conversely, one could argue that the grand coalition should have an advantage, since collusion is an illegitimate activity. In this paper, we adopt a view which, surprisingly, has not been very prominent in the theory of contracts. We assume all coalitions, including the grand coalition, have access to the *same* contracting technology. An *original contract* regulates the relationship between the three agents. The original contract is supervised by an impartial *original judge*, who collects messages from the agents and orders both "real" actions and transfers of money. The original judge may not be an actual person, but a computer program that collects inputs and selects an output. Collusion consists of a subcoalition secretly signing a *collusive contract*, in the spirit of Laffont and Martimort [13]. In order to maintain symmetry between the original contract and the collusive contract, the latter is supervised by a *collusive judge* who is similar to the original judge (and so may not be a physical person).

It is important to specify precisely what a judge can observe, because things he can observe are "verifiable information" for the contract he supervises. If the original contract implements a real action, such as a trade of commodities, then this real action is publicly observed. Therefore, it is verifiable by a collusive judge, which facilitates collusion. But any mechanism may specify *secret messages* and *secret cash transfers* which cannot be verified by outside judges. On the one hand, this assumption seems reasonable. On the other hand, if the collusive agreement cannot be kept secret, then by definition it is verifiable information in the original contract, and the original judge can punish any colluding agent. Therefore, to make the problem non-trivial,

collusion must take place behind closed doors, unobserved by outsiders. Since all coalitions have access to the same contracting technology, it must also be possible for the original mechanism to be played out behind closed doors.[1] In effect, we assume no judge has any cameras or microphones installed in the any other judge's court room. (Alternatively, except for the real action, the inputs and outputs to one computer program cannot be direct inputs into the other.) Of course, even if a collusive judge cannot directly observe what goes on in the original judge's court room, he may try to infer it by observing which real action the original judge orders, or by asking the colluding agents to report what they know.

In the implementation literature, "full implementation" requires that there are no equilibrium outcomes other the desired outcome. If, on the other hand, the desired outcome is only one of many possible equilibrium outcomes, then the mechanism "weakly implements" the desired outcome. When the original contract is designed, the main difficulty is to ensure the existence of a *collusion-free* equilibrium which achieves an efficient outcome. Consequently, we focus on the problem of weak implementation. Once this problem is solved, full implementation is fairly easy to accomplish. Indeed, there are many ways to rule out the possibility of "bad" (collusive) equilibria. For example, suppose we want to rule out a "bad" equilibrium where the builder and the third party collude against the architect. In such a collusive equilibrium, the architect *knows* that he is colluded against and that his payoff is likely to be low. (In general, each player knows which equilibrium is played.) But it is easy to include a message ("blowing a whistle") that the architect prefers to send if (and only if) he knows he is colluded against, thereby destroying the collusive equilibrium.

We model collusion in a non-cooperative way. The agents propose and accept collusive contracts as part of an extensive form game. In order to achieve weak implementation of the first best outcome, we design an original mechanism which has a collusion-free perfect Bayesian equilibrium (PBE). No collusion occurs along the equilibrium path. To support this collusion-free PBE, we must ensure that, for any collusive coalition and any collusive mechanism, there exists a continuation equilibrium[2] which would be bad for at least one of the colluding agents. Suppose the builder and the third party have access to a collusive mechanism which induces multiple continuation equilibria, one of which is bad for the builder. In the collusion-free PBE, if the third party were to propose this collusive mechanism, the builder would reject it, anticipating the bad outcome. Of course, the third party can propose

[1] Hart and Moore [9] argue that it might be impossible to outlaw collusion even if the side-contract is publicly observed, because the side-contract might be so complicated that an outsider cannot understand its true (collusive) meaning. In that case, we would simply assume that the original judge's court hearings are also incomprehensible to outsiders, and our results go through.

[2] Since collusion is secret, outsiders do not know that a coalition has formed. By a continuation equilibrium we mean the actions the agents plan to take, conditional on what they know.

any collusive mechanism whatsoever. If he proposes a collusive mechanism which induces a *unique* continuation equilibrium which is good for *both* him and the builder, then the builder would accept it, and the collusion-free PBE would collapse. In other words, collusion is a threat if and only if a collusive mechanism *fully* implements an outcome that is good for all the colluding agents.[3] We show that no such collusive mechanism exists if the original contract is well designed. To see the intuition, suppose that the builder and the third party agree to tell the collusive judge what happened when the original mechanism was played out. *Full* implementation of the collusive agreement will be difficult, because (with quasi-linear utility functions) past messages and transfers are not "payoff relevant".

To make our possibility result as strong as possible, we make a number of assumptions that facilitate collusion. First, we assume collusion is secret, so the original judge cannot simply outlaw it. Second, the collusive judge can observe *real* outcomes, such as the trade that is specified by the original mechanism or the output of the team. Third, we assume there can only be one round of collusion. Thus, a colluding coalition does not have to worry about making the collusive agreement robust against further rounds of collusion. Fourth, we assume the collusive coalition can use inefficient punishments, such as destroying wealth. Notice that the third and fourth assumptions imply that subcoalitions will in fact enjoy a contracting advantage over the grand coalition. Since these assumptions facilitate collusion, they *strengthen* our possibility result (namely, that the first best can be implemented by a collusion-free original mechanism). In addition, they simplify the analysis.[4]

In the buyer-seller model, it is possible to implement the first best outcome using a very simple original mechanism which always recommends the same real action (i.e., the same trade), regardless of the messages. A collusive judge will be unable to figure out which messages were sent in the original mechanism, and it is impossible for colluding parties to (uniquely) implement a collusive outcome of the form "if the third party receives a fine, he will share it with his colluding partner." For any collusive mechanism, there will exist a continuation equilibrium where the third party makes the *same* transfer to his colluding partner, regardless of what happened in the original mechanism. But this only adds or subtracts a constant from the payoffs and

[3] This does not mean that we give some kind of advantage to the grand coalition by requiring weak implementation for them, but full implementation for subcoalitions. The fact that a collusion-free PBE can be constructed if no subcoalition can achieve full implementation is not an assumption, it is an implication of the definition of PBE. Although we focus on weak implementation for the grand coalition, as we have explained it is easy to adjust the original mechanism to ensure full implementation.

[4] Without the third and fourth assumptions, collusion between two parties might benefit from an outside budget-breaker. This would require us to introduce a "fourth party" who breaks the budget for colluding two-agent coalitions. But this opens up further possibilities of collusion. The analysis would be very complicated.

does not affect marginal incentives within the original mechanism. Hence, it does not threaten the collusion-free equilibrium.

In the model of moral hazard in teams, the total output is a "real" amount of goods produced. This is verifiable for any judge. Unlike Holmström [12], we assume collusion is possible.[5] A collusive contract can implement outcomes where the third party makes a transfer to a team member who is colluding with him, as a function of the (verifiable) output of the team.[6] Such a collusive contract does affect the team member's incentives in the original mechanism, and may potentially upset a collusion-free equilibrium. Indeed, we show that if the original contract does not specify a message game, then a third party is of no value. However, if message games are allowed, then a third party is valuable. We construct an original mechanism with a "whistle blowing" clause, where the third party reveals what he knows about collusion, and we make sure the whistle-blower is safe from punishment by the collusive judge. Whatever mechanism the colluding agents design, there is always a continuation equilibrium where one of them "blows the whistle" in front of the original judge, who then punishes the other colluding agent. Therefore, the one who will be punished does not want to collude. Notice that a message game is necessary for weak implementation of the first best, even though the agents have no information about each other's effort levels. In the presence of collusion, the revelation principle cannot be interpreted to mean that the agents should simply tell the original judge what they know about effort levels. The original judge should also try to elicit information about any collusive agreements that the agents have entered into.

Our buyer-seller model is similar to Segal and Whinston [20], and encompasses models which have used Maskin and Moore's [14] implementation with renegotiation paradigm to provide foundations for "incomplete contracts". For example, in Che and Hausch [3], the decision the traders face is what quantity to trade and one agent's investment is allowed to affect the other's payoff (cooperative investments). The first-best cannot always be achieved but the second-best can be implemented without writing an explicit contract at all. In Segal [19], n possible goods can be traded and each agent's investment only affects his own payoff.[7] Under some assumptions, as n becomes large, the first-best cannot be achieved, and the second-best can be

[5] Eswaran and Kotwal [6] introduce collusion into Holmström's team model. Brusco [2] looks at collusion in a model where the team members can observe each others' effort levels.

[6] In this model, a "third party" (i.e., an outsider who is not a team member) can be valuable even if the team has more than two members. This is because individual effort is unobserved, so implementation is difficult even with three or more agents. It is only with three or more *completely informed* agents that implementation is easily achieved by message games (e.g., Sjöström [21]). However, to maintain symmetry with the buyer-seller model, we focus on a two-agent team.

[7] See also Maskin and Tirole [17] and Hart and Moore [11].

implemented without any contract at all. We show that implementation of the first-best can be achieved by the introduction of a third party, even if agents can collude and renegotiate inefficient outcomes. This puts into question the foundations for "incomplete contracts".

In some models of bilateral hold-up with renegotiation, the first best can be implemented even without a third party. Che and Sákovics [4] consider a dynamic model where the timing of investment is endogenous. They show that the hold-up problem can be resolved, even with no contract, if the buyer and seller are sufficiently patient. Aghion, Dewatripont and Rey [1], Nöldeke and Schmidt [18] and Edlin and Reichelstein [5] also found positive results for the bilateral case. These results complement ours by establishing conditions where a third party is not needed. They reinforce our conclusion that the first-best can be implemented in a broad range of hold-up models.

The rest of the paper has two parts. Section 2 contains the buyer-seller model with a third party. Section 3 studies the moral hazard in teams model.

2. THE BUYER-SELLER MODEL

2.1. The Buyer-Seller Relationship. There is a buyer B and a seller S. Let $b \geq 0$ denote a relationship-specific investment made by the buyer and let $s \geq 0$ denote a relationship-specific investment made by the seller. Let $\omega \in \Omega$ denote a random variable which is realized after investments are made. The buyer's realized cost of making investment b is $\varphi_B(b, \omega)$, and the seller's realized cost of making investment s is $\varphi_S(s, \omega)$. The vector $\theta = (b, s, \omega)$ is the *state of the world*. We make the standard assumption that θ is observed by B and S, but by no-one else.

Trade between the buyer and seller is represented by a set of possible real actions denoted X. Specifically, a real action $x \in X$ may specify what kind of good (and how much of it) the seller delivers to the buyer. The buyer's gross value from the trade is denoted $v(x, b, s, \omega)$. The seller's cost from the trade is $c(x, b, s, \omega)$. This formulation is quite general. For example, it allows the possibility of "cooperative" investments. There is a "null outcome" $x_\emptyset \in X$ which we interpret as "no trade". In additions to the "real" actions in X, monetary transfers can be made. Utility functions are quasi-linear in money. Thus, for example, if the buyer receives a monetary transfer t_B then his final payoff is $v(x, b, s, \omega) - \varphi_B(b, \omega) + t_B$.

The *ex post surplus* is $v(x, \theta) - c(x, \theta)$. Let $x^*(\theta) \in X$ be the real action (assumed unique) which maximizes the expected surplus in state θ. That is,

$$x^*(\theta) \equiv \arg\max_{x \in X} \{v(x, \theta) - c(x, \theta)\}$$

Define

$$v^*(\theta) \equiv v(x^*(\theta), \theta)$$

and

$$c^*(\theta) \equiv c(x^*(\theta), \theta)$$

The *maximized ex post surplus* is

$$\Sigma^*(\theta) \equiv \max_{x \in X} \{v(x, \theta) - c(x, \theta)\} = v^*(\theta) - c^*(\theta)$$

The *first best investment levels* (b^*, s^*) maximize the expected value of $\Sigma^*(b, s, \omega) - \varphi_B(b, \omega) - \varphi_S(s, \omega)$, where the expectation is with respect to the random variable ω. That is,

$$(b^*, s^*) \equiv \arg\max_{(b,s) \geq 0} E_\omega \{\Sigma^*(b, s, \omega) - \varphi_B(b, \omega) - \varphi_S(s, \omega)\}$$

We assume for simplicity that there is a unique first-best pair (b^*, s^*). The *first-best solution* to the contracting problem is for the buyer and seller to make investments (b^*, s^*), and for every $\omega \in \Omega$, to take the real decision $x^*(b^*, s^*, \omega)$. If the first-best solution if implemented with transfers t_B and t_S, then the buyer's expected payoff is

$$B^*(t_B) \equiv E_\omega\{v^*(b^*, s^*, \omega) - \varphi_B(b^*, \omega)\} + t_B$$

The seller's expected payoff is

$$S^*(t_S) \equiv E_\omega \{-c^*(b^*, s^*, \omega) - \varphi_S(s^*, \omega)\} + t_S$$

2.2. The Third Party. A third party T may be invited to play the role of *budget-breaker* in the buyer-seller relationship. Thus, there are three players: B, S, and T. The third party cares only about money (not about x or θ) and his payoff is linear in wealth. He does not observe θ.

In order to facilitate collusion, a colluding strict subset of $\{B, S, T\}$ will be allowed to use inefficient punishments (such as destroying wealth). Because of this assumption, there is no need for a colluding pair to enlist the services of a "fourth party". Indeed, in economic environments such as the one we are considering, if inefficient outcomes cannot be renegotiated then two-person implementation is no more difficult than three-person implementation (e.g., Maskin and Sjöström [15]). There is no need for a budget-breaker if the budget does not have to balance, so a "fourth party" would be useless. But in the original contract we do not allow any kind of inefficiency, including the destruction of wealth, so the third party is not useless.

2.3. Time Line. The relationship between B, S and T is governed by an *original contract* which specifies an *original mechanism* Γ_0. A mechanism is a *message game*, which specifies message spaces and an outcome function which maps messages into outcomes.[8] We do not model the bargaining process which produces Γ_0. We simply

[8] We restrict attention to normal form mechanisms. Allowing the agents to send messages sequentially would not change any results.

assume that at the beginning of the game, Γ_0 has already been determined. The extensive form game $G(\Gamma_0)$ induced by Γ_0 is described by the following time-line.

At time 0, there is a coalition-formation game with two stages. In the first stage, T can make a collusive proposal to B or S (but not to both).[9] A proposal to player $j \in \{B, S\}$ is an invitation to form a two-player coalition $C = \{j, T\}$. The proposal specifies a set of "collusive mechanisms" $\{\Gamma_C(x)\}_{x \in X}$.[10] As the notation suggests, we allow the collusive mechanism to depend directly on x, the action implemented by the original mechanism, because x is verifiable. At time 0, the colluding parties do not know what real action will be implemented by Γ_0, so they write a contingent collusive contract of the form "if Γ_0 produces the outcome x, then the colluding coalition will play message game $\Gamma_C(x)$ at time 4".

If T does not make any collusive proposal in the first stage, then we bypass the second stage and proceed to time 1. If T makes a proposal to player $j \in \{B, S\}$ in the first stage, then in the second stage player j responds to T by either accepting or rejecting the proposal. If player j accepts then coalition $\{j, T\}$ has formed and the collusive proposal is in force. If j rejects (or no proposal was made in stage one), then no coalition is formed.

Collusion is done secretly. Thus, if T proposes a coalition $\{B, T\}$, then S is not informed about this (neither is S informed about B's decision to accept or reject the proposal). Similarly, B is never told about any collusion between T and S. Consequently, if a two-person coalition forms, then the party who is left out is not informed about this and cannot react to it in any way.

At time 1, the buyer and the seller make investments $b \geq 0$ and $s \geq 0$, respectively. The buyer observes the seller's investment s, and the seller observes the buyer's investment b, but no-one else observes b or s.

At time 2, the random variable $\omega \in \Omega$ is realized. The realization of ω is observed by the buyer and the seller, but by no-one else.

At time 3, the original mechanism Γ_0 is played out among the original parties. The mechanism specifies a message space M_i for each player $i \in \{B, S, T\}$. For each message profile $m \in M_B \times M_S \times M_T$, the mechanism produces an outcome of the

[9] It is evident that collusion between B and S will not be a problem. Therefore, allowing B and S to make proposals would not change anything.

[10] The only restriction we put on the set of possible collusive mechanisms is that they be regular enough that a continuation equilibrium exists. If the strategy space in a collusive mechanism is an open set, for example, no continuation equilibrium may exist. Presumably, the collusive judge would not tolerate it. Formally, we assume T can only propose collusive mechanisms with the best-response property: each player will aways have a best response to any strategy chosen by his opponents.

form $(x(m), t(m))$. Here $x(m) \in X$ is a real action ordered by the original judge, and $t(m) = (t_B(m), t_S(m), t_T(m)) \in \mathbf{R}^3$ is a vector of monetary payments, where $t_i(m)$ is a transfer to player i. We require $t_B(m) + t_S(m) + t_T(m) = 0$ for all m.[11]

At time 4, nothing happens if no coalition was formed at time 0. However, if a coalition C was formed, then the collusive mechanism $\Gamma_C(x)$ is played out among the colluding parties (where $x = x(m)$ is the real decision determined by the messages m at time 3). The collusive mechanism $\Gamma_C(x)$ specifies a message space $M_i^C(x)$ for each player $i \in C$. For each message profile $m^C \in \times_{i \in C} M_i^C(x)$, $\Gamma_C(x)$ produces an outcome $(t_i^C(m^C))_{i \in C} \in \mathbf{R}^{|C|}$. Here $t_i^C = t_i^C(m^C)$ is a monetary payment to agent $i \in C$. We require $\sum_{i \in C} t_i^C \leq 0$. The collusive mechanism $\Gamma_C(x)$ cannot specify a different real action than the original mechanism. Any attempt to overrule the original contract by choosing a different x would constitute a violation of the original contract, which we assume can be ruled out by the original judge. Indeed, since the physical transaction x directly involves B and S (but not T), a coalition that excludes either B or S clearly cannot have any right to choose x.[12] However, the collusive mechanism can safely specify cash transfers among the colluding parties, because these are not publicly observable.

At time 5, B and S may renegotiate the decision $x = x(m)$ produced by Γ_0 at time 3. The renegotiation takes place in secret and cannot be observed by anyone except B and S. With probability λ_B, the buyer makes a take-it-or leave-it proposal, consisting of a new decision $x^R \in X$ and a pair of transfers (t_B^R, t_S^R) (which are added to any previous transfers the players have received). We require $t_B^R + t_S^R = 0$. If the seller accepts, the proposal is implemented. If the seller rejects, there is no renegotiation and the game ends. With probability $\lambda_S = 1 - \lambda_B$, it is the seller who makes a take-it-or leave-it proposal. If the buyer accepts, the proposal is implemented. If the buyer rejects, there is no renegotiation and the game ends. Our results are not sensitive to the precise specification of the renegotiation game. Any reasonable specification (e.g., alternating-offer bargaining) would lead to the same results.

[11] Thus, we do not give the original judge the power to destroy wealth.

[12] If the real action ordered by the collusive judge is publicly observable, then by definition is it verifiable and can be ruled out by the original contract. On the other hand, to assume a collusive judge could "secretly order" a real action would have absurd consequences. To be specific, suppose the judge supervising a secret collusive agreement between B and T has the power to order a "secret trade" between B and S. Then S might receive an order to secretly deliver some goods to B, signed by a judge he never heard of. We assume that, since S is not a part of the collusive agreement - indeed it is directed against him - he has no obligation to obey this order. That is, a judge has no jurisdiction over agents other than those who signed the agreement he supervises. Otherwise, B and T could sign an absurd agreement which would force S to hand over everything he owns.

2.4. The Renegotiated Outcome. The game $G(\Gamma_0)$ is solved backwards. Suppose time 5 has been reached and consider the continuation equilibrium of the renegotiation stage. The true state of the world $\theta = (b, s, \omega)$ is known to B and S. The mechanism Γ_0 has recommended the real decision $x = x(m)$ at time 3. Player i receives a transfer $t_i = t_i(m)$ at time 3. If a coalition C formed at time 0, then player $i \in C$ also receives a transfer t_i^C at stage 4. Notationally, if $i \notin C$ then set $t_i^C \equiv 0$. Let $\hat{t}_i$ denote the sum of the transfers,

$$\hat{t}_i = t_i + t_i^C \tag{1}$$

for $i \in \{B, S, T\}$.

In the continuation equilibrium, the renegotiated outcome x^R will maximize ex post surplus, i.e., $x^R = x^*(\theta)$. Whichever party makes the take-it-or-leave-it offer will appropriate all the surplus. If B makes the proposal, he will make sure that S is indifferent between accepting and rejecting the proposal. In order to convince S to switch from x to $x^*(\theta)$, S will be compensated by the amount $t_S^R = c^*(\theta) - c(x, \theta)$. Conversely, if S makes the proposal, then B will be compensated by the amount $t_B^R = v(x, \theta) - v^*(\theta)$.

Given state θ, real decision x as recommended by Γ_0, and the pair of transfers $(\hat{t}_B, \hat{t}_S)$, we can now calculate the buyer's expected payoff, taking renegotiation into account, but not including the cost of the investment. It is

$$\begin{aligned} & u_B(x, \hat{t}_B, \theta) = v^*(\theta) + \hat{t}_B - \lambda_B \left[c^*(\theta) - c(x, \theta)\right] + \lambda_S \left[v(x, \theta) - v^*(\theta)\right] \qquad (2) \\ = \;& \lambda_B \Sigma^*(\theta) + \hat{t}_B + \lambda_B c(x, \theta) + \lambda_S v(x, \theta) \end{aligned}$$

Similarly, the seller's expected payoff is

$$u_S(x, \hat{t}_S, \theta) = \lambda_S \Sigma^*(\theta) + \hat{t}_S - \lambda_B c(x, \theta) - \lambda_S v(x, \theta) \tag{3}$$

Since we know what will happen at time 5, we will suppress the renegotiation stage in what follows. Thus, if there is no collusion and the message-game form Γ_0 produces the outcome $(x(m), t_B(m), t_S(m), t_T(m))$ at time 4, then the buyer's final payoff will be $u_B(x(m), t_B(m), \theta)$, as defined by (2). The seller's final payoff will be $u_S(x(m), t_S(m), \theta)$, as defined by (3). The third party's payoff will be $t_T(m)$. If there is collusion, then the collusive transfers $(t_i^C)_{i \in C}$ are added to the payoffs in the obvious way, according to (1).

2.5. Participation Constraints. The buyer and seller may have the option of not trading. To formalize this, let Γ_0^* denote a "null" mechanism which simply recommends the outcome $x_\emptyset$ and no transfers (there are no messages). Of course, the outcome $x_\emptyset$ will be renegotiated at time 5. The payoffs in state (b, s, ω) will be

$$u_B(x_\emptyset, 0, (b, s, \omega)) - \varphi_B(b, \omega) \tag{4}$$

for the buyer and

$$u_S(x_\emptyset, 0, (b, s, \omega)) - \varphi_S(s, \omega) \tag{5}$$

for the seller, using the definitions in (2) and (3). Under the null contract, the buyer will set b to maximize the expectation of (4) with respect to ω, taking s as given. The seller will set s to maximize the expectation of (5) with respect to ω, taking b as given. Let $\hat{b}$ and $\hat{s}$ denote the equilibrium investments. In general these investments will not be at the efficient level, $(\hat{b}, \hat{s}) \neq (b^*, s^*)$. The expected payoffs from the equilibrium induced by the null mechanism Γ_0^* are

$$B_\emptyset \equiv E_\omega \left\{ u_B(x_\emptyset, 0, (\hat{b}, \hat{s}, \omega)) - \varphi_B(\hat{b}, \omega) \right\} \tag{6}$$

and

$$S_\emptyset \equiv E_\omega \left\{ u_S(x_\emptyset, 0, (\hat{b}, \hat{s}, \omega)) - \varphi_S(\hat{s}, \omega) \right\} \tag{7}$$

Of course, with the null contract, the third party plays no role and gets zero payoff.

The participation constraints will ensure that both B and S are better off than under the null contract. It is useful to define a pair of transfers (t_B^*, t_S^*), where $t_B^* = -t_S^*$, such that

$$B^*(t_B^*) \equiv E_\omega\{u_B(x_\emptyset, t_B^*, \theta) - \varphi_B(b^*, \omega)\} \geq B_\varnothing \tag{8}$$

and

$$S^*(t_S^*) \equiv E_\omega\{u_S(x_\emptyset, t_S^*, \theta) - \varphi_S(s^*, \omega)\} \geq S_\varnothing \tag{9}$$

Suppose the buyer and seller make the first-best investments b^* and s^*, respectively, and in every state $x_\emptyset$ is implemented and transfers (t_B^*, t_S^*) are made. Renegotiation will take no-trade to the first-best decision $x^*(\theta)$ in every state, and (8) and (9) guarantee that the buyer's and the seller's participation constraints are satisfied. That is, they are better off than they would be under the null contract.

2.6. Discussion. Here we further discuss our assumptions about observability and verifiability. First, we make the standard assumption that the state is observed by the buyer and the seller, but unobservable (and hence unverifiable) to outsiders, including the third party and the judges.

Second, we assume coalitions can form secretly. Thus, the original contract cannot simply outlaw coalition formation. Maskin and Tirole [17] suggest that the original contract might reward any agent who produces evidence of a collusive contract (and the original contract will punish the other colluding parties). We assume hard evidence about collusive deals is impossible to produce. However, any member of a collusive coalition will have "soft" information about collusion, and can be asked to reveal it ("whistle-blowing"). Thus, at time 3 a "revelation mechanism" should not

just collect reports about the state of the world, but also about collusive deals. Of course, the agents must be given an incentive to tell the truth. (It turns out that whistle-blowing is required for efficiency in the team model, but not in the buyer-seller model).

Third, what goes on behind the closed doors of a judge (messages and cash payments) cannot be observed by outsiders. This assumption is needed to make collusion possible. Indeed, if the original judge could observe the collusive proceedings, then collusion would be verifiable information, and so the original judge could outlaw it. By symmetry, the collusive judge cannot observe the original mechanism being played out either.[13] Consequently, the outcome of a mechanism cannot depend *directly* on what happens in *another* mechanism (the exception is the real action x, which is verifiable). In addition to making the problem non-trivial, the assumption of unobserved court proceedings is fairly plausible.[14]

Fourth, we assume that the real decision x produced by Γ_0 at time 3 is publicly observable, hence verifiable by any judge. The real decision x has a physical manifestation outside the court room, so unlike messages and monetary transfers, it is impossible to keep x secret. For example, the original judge may order the seller to deliver a certain quantity of goods to the buyer. This physical action cannot take place in secret. This assumption helps the agents collude, because the colluding parties can make their agreement conditional on x.

Fifth, we make the standard assumption that renegotiation at time 5 is unverifiable. If renegotiation is verifiable, then the original mechanism can prescribe that large payments be made by B and S to T should the final decision x' differ from the decision x prescribed by the original contract. Maskin and Tirole [17] suggest that the original contract might reward any agent who produces evidence of renegotiation (and the contract will punish the others who participated in the renegotiation).[15] With such a scheme, even if renegotiation occurs, some of the surplus generated by

[13] If the collusive judge has a microphone installed in the original judge's court room, but the original judge does not have any microphone installed in the collusive judge's court room, then the colluding coalition has a more powerful contracting technology than the grand coalition. But the purpose of our paper is to see what happens if all coalitions have access to the same technology.

[14] If agent i gives cash to the judge who transfers it to agent j, neither agent will have any proof that the transaction took place. The judge does not give out any receipts (he is incorruptible so no receipts are necessary). A bank statement that cash has been withdrawn by agent i from his account does not reveal what happened to the money. Even if the mechanism is a computer program, it can automatically open a new bank account in agent j's name and deposit money in it. It would be impossible for agent j to prove to that he did not receive money in this way. Notice that this scheme makes it possible to secretly reward whistle blowers.

[15] A difficult situation occurs if a new renegotiated contract surfaces which contradicts the original contract. Suppose the new contract is signed by B, T and S and contains a clause invalidating Γ_0. It is not clear which contract would in fact be enforced. If the new contract has precedence, then renegotiation cannot be eliminated by Γ_0 in the way Maskin and Tirole suggest.

it might go to T, so renegotiation might be costly for B and S. By adding more and more parties to the contract, the share of the surplus going to B and S might be lowered even further. However, we assume evidence of renegotiation is impossible to produce, so renegotiation is impossible to rule out in the original contract.[16] This assumption makes it more difficult to design Γ_0 to implement efficient outcomes.

Given our assumptions, a collusive contract signed by a coalition C can specify transfers as a function of the decision x ordered by the original judge and the messages sent in the collusive message game. This might indirectly influence which decision x is implemented by the original contract. For example, a badly designed original contract might specify that if B says the quality of the good is low then the outcome is $x_\emptyset$ and S pays a fine to T, but otherwise they trade a positive amount. Now B and T can secretly agree that B should always report that the quality is low and split the fine with T. The collusive contract can enforce this by specifying that, if the outcome produced by the original mechanism is anything else than $x_\emptyset$, then B pays a large fine to T. But a well designed original contract wouldn't reveal the messages in this blatant way. A collusive message game might still be used to elicit information about the messages sent in the original mechanism. That is, the collusive judge could ask B and T about what B told the original judge.

In sum, we have chosen assumptions in such a way that renegotiation and collusion are facilitated as much as possible, subject to the constraint that all coalitions should use the same contracting technology. Nevertheless, we show that efficient allocations can be implemented using a well-designed Γ_0.

2.7. Implementation. We will design an original mechanism Γ_0 and construct a collusion-free equilibrium of $G(\Gamma_0)$ (where no coalition forms along the equilibrium path), which produces the first-best solution. In this section, we will not worry about the possible existence of other, non-optimal, equilibria. Thus, this section deals with "weak" implementation of the first best.

In order to support the collusion-free equilibrium, T should not have any incentive to make a collusive proposal at time 0. To ensure this, we need to show that for any colluding coalition, there exists a continuation equilibrium which is bad for at least one of the colluding agents. That is, for any possible collusive proposal T could make to player $j \in \{B, S\}$, either the proposal makes player j worse off, so he will reject it, or T is made worse off, so the he does not want to make the proposal in the first place.

We now define the original mechanism Γ_0 which is played out at time 3. This particular mechanism will be called the *secret message mechanism*. In the secret

[16] Another reason to allow renegotiation is that B and S may want to trade again in the future. Ruling out future transactions might be inefficient if not impossible (B and S may use intermediaries to trade).

message mechanism, the buyer and the seller simultaneously announce the state. Formally, player $i \in \{B, S\}$ sends a message θ^i from the message space $M_i \equiv \Theta$. The third party sends no message. To ensure that the participation constraints are satisfied, the pair of transfers (t_B^*, t_S^*) satisfy (8) and (9), with $t_B^* = -t_S^*$. The outcome function is defined as follows.

Rule 1. If $\theta^B = \theta^S = \theta$, then the real decision is $x(m) = x_\emptyset$, and transfers are determined as follows. If $\theta = (b^*, s^*, \omega)$, then the buyer pays t_S^* to the seller. If $\theta = (b, s^*, \omega)$ with $b \neq b^*$ then the buyer pays F^1 to the seller. If $\theta = (b^*, s, \omega)$ with $s \neq s^*$, then the seller pays F^1 to the buyer. If $\theta = (b, s, \omega)$ with $b \neq b^*$ and $s \neq s^*$ then no transfers are made.

Rule 2. If $\theta^B \neq \theta^S$ then $x(m) = x_\emptyset$. The buyer and seller each pay F^2 to the third party.

Although the messages are observed by B, S and T, they are not verifiable by outsiders. The no-trade outcome is always implemented to avoid signaling the message profile indirectly. We will show that a collusive mechanism cannot elicit information about the original messages, so there is no way to collude profitably. The collusion-free equilibrium is first-best, because the no-trade outcome is renegotiated to the efficient decision in every state, and transfers are designed to give efficient incentives.

Theorem 1. *We can choose F^1 and F^2 so that the game $G(\Gamma_0)$ has a collusion-free perfect Bayesian equilibrium which produces the first-best outcome. Transfers (t_B^*, t_S^*) are implemented by the mechanism in every state, so the participation constraints are satisfied.*

To prove the theorem, we construct a perfect Bayesian equilibrium of $G(\Gamma_0)$ as follows. At time 0, no collusive proposal is made. If the players have not joined any collusive coalition at time 0, then they play as follows from time 1 on. The buyer and seller invest at the first best-level at time 1, and at time 3 they tell the truth (in all states of the world). If at time 3, either the buyer or the seller deviates and lies about the state, then they incur the fine F^2 by Rule 2. If F^2 is large enough, neither the buyer nor the seller has an incentive to deviate from truth-telling. Furthermore, if F^1 is sufficiently big, then Rule 1 implies that both agents prefer to choose the first-best investment levels, anticipating that the truth is revealed at time 3. >From this it follows that, if there is no collusion at time 0, then the proposed strategies are sequentially rational from time 1 on. The outcome is first best by construction, and the equilibrium payoffs are $B^*(t_B^*)$ for the buyer, $S^*(t_S^*)$ for the seller, and 0 for the third party.

It remains to specify behavior after a time 0 deviation where T makes a collusive proposal. To support the equilibrium, such a deviation should not be profitable. To

be specific, suppose T makes a collusive proposal to B (the argument concerning a proposal to S will be exactly the same). We construct the strategies so that if B accepts and coalition $C = \{B, T\}$ forms, then either B or T will get no more than their equilibrium payoff. If T gets no more than zero, it is certainly not profitable for him to propose the coalition. If B gets no more than $B^*(t_B^*)$ from joining the coalition, then we stipulate that his equilibrium strategy is to reject the proposal, and this behavior is certainly sequentially rational. Knowing that B will reject, it is again not profitable for T to make the proposal.

So suppose B accepts. The coalition $C = \{B, T\}$ forms and a collusive proposal $\{\Gamma_C(x)\}_{x \in X}$ is in force. S is unaware of the collusion and will play as described above. The equilibrium strategies must specify how B and T will behave in $G(\Gamma_0)$ after they have formed a coalition. Consider a pair of collusive messages $(m_B^C, m_T^C) \in M_B^C(x_\emptyset) \times M_T^C(x_\emptyset)$ such that

$$t_B^C(m_B^C, m_T^C) \geq t_B^C(m_B, m_T^C)$$

for all $m_B \in M_B^C(x_\emptyset)$, and

$$t_T^C(m_B^C, m_T^C) \geq t_T^C(m_B^C, m_T)$$

for all $m_T \in M_T^C(x_\emptyset)$. That is, (m_B^C, m_T^C) would be a Nash equilibrium of a game where only the collusive transfers matter. Some such pair must exist (we can allow mixed strategies), because T is not allowed to propose a badly behaved collusive mechanism that causes an existence problem. Now we let the equilibrium strategies for $G(\Gamma_0)$ specify that, when C has formed, B and T choose this particular pair (m_B^C, m_T^C) *regardless* of what else has happened in the game before time 4. This can be done because nothing that happens before time 4 can change the strategic incentives in $\Gamma_C(x_\emptyset)$, which are always just to maximize one's collusive sidepayment. So (m_B^C, m_T^C) will be part of a continuation equilibrium, following any history. With these strategies, B's investment and the message he sends at time 3 do not affect the collusive transfers, so B maximizes his payoff by making the first-best investment and telling the truth at time 3. Rule 1 will apply, and S will get a payoff $S^*(t_S^*)$. But then, either B gets no more than $B^*(t_B^*)$ or T gets no more than zero. As argued above, this implies that T does not gain by making the proposal. This completes the proof of Theorem 1.

The buyer and the third party would jointly benefit if they could enforce the following side contract: "The third party pays the buyer $2F_2 - \varepsilon$ at time 4 if and only if the buyer contradicted the seller at time 3." But this is not an enforceable side contract, because messages in Γ_0 are not verifiable by the collusive judge, and the real action is always $x_\emptyset$. Moreover, using a collusive message game to elicit information about messages sent in Γ_0 won't work. With quasi-linear utilities, previous

transfers are payoff-irrelevant, so there is always an "uninformative" continuation equilibrium where the time 4 messages are independent of what happened at time 3. But then, B may as well tell the truth at time 3, to avoid paying an extra fine to T.

We offer two further comments on Theorem 1.

First, like most of the literature we assume B and S are risk-neutral. However, suppose instead that they have strictly concave von Neumann-Morgenstern utility functions. If the degree of risk aversion varies with wealth, and if a colluding coalition can implement lotteries at time 4, then previous transfers become payoff relevant. At time 4, a collusive lottery mechanism might extract indirect information about previous transfers, and hence enforce a collusive scheme. However, if such lottery mechanisms are practical for a colluding coalition, then they can also be put into the original contract, if we maintain the hypothesis that the contracting technology for any coalition should be the same. In this case, it follows from Maskin and Moore [14] that the buyer and seller can implement the first best even without the help of a third party. On the other hand, Hart and Moore [11] argued that lottery mechanisms are impractical. In this case, the first-best can be implemented with the help of a third party, by using the secret message mechanism described above, because a colluding coalition cannot use a lottery mechanism to extract information about what happened at time 3.

Second, we have assumed that the set of possible real actions X is describable ex ante. If parts of it are indescribable, it is impossible to write an original contract that fully identifies which action to implement. Indeed the typical assumption in the incomplete contracts literature is that some actions, such as asset ownership, are always describable. Others, such as which object to trade, are indescribable ex ante but describable ex post (e.g., Grossman and Hart [7], Hart and Moore [10] and Hart [8]). The indescribability may be pertinent as it may be optimal to trade different objects in different states of the world. However, as Maskin and Tirole [16] argued, there is a tension between the assumption that certain actions are indescribable ex ante and the assumption that agents are able to perform dynamic programming. Maskin and Tirole's [16] Theorem 4 shows that a contract that is implementable (with renegotiation) when actions are describable is also implementable (with renegotiation) when they are indescribable. In this sense, indescribability is irrelevant and the only binding constraints are those imposed by renegotiation. Thus, even if actions are indescribable, Maskin and Tirole's [16] irrelevance theorem together with the results of our paper imply that a third party contract can implement the first best.

2.8. Full Implementation. There are various ways to make sure that *all* perfect Bayesian equilibria produce the first-best outcome. One simple way is to amend the time line in Section 2.3 by allowing B and S to send messages at the very beginning

of the game, just before time 0. We call this time -1. The announcements at time -1 will decide whether the mechanism played at time 3 will be the secret message mechanism described in Section 2.7, or the null mechanism Γ_0^* described in Section 2.5 (recall that Γ_0^* has no messages, and always recommends $x_\emptyset$ and no transfers). The other parts of the time line, such as the secret collusion at time 0, are unchanged.

Specifically, the *augmented secret message mechanism* works as follows. At time -1, B and S simultaneously announce non-negative integers. These announcements may be taken to be publicly observed and verifiable. There are two cases.

Case 1. Suppose someone announces a strictly positive integer at time -1. Then there are no messages and no transfers at time 3, and the outcome produced in period 3 is $x_\emptyset$. That is, they play the null mechanism at time 3. However, the player who announces the highest integer receives a cash payment from his trading partner at time -1.[17] If the player with the highest integer is B, then S must pay B the amount

$$\tau_B = B^*(t_B^*) - B_\emptyset \tag{10}$$

That is, the transfer equal to the difference between B's first-best payoff and the payoff from playing the null mechanism. Similarly, if the player with the highest integer is S, then B must pay S the amount

$$\tau_S = S^*(t_S^*) - S_\emptyset \tag{11}$$

Thus, in case 1, the payoffs will be the same as under the null mechanism, with the time -1 transfers added on. Notice that in this case, collusion is moot.

Case 2. Suppose both B and S say zero at time -1. Then the game unfolds just as described in Section 2.7. Thus, at time 3 the secret message mechanism described in Section 2.7 is operated. That is, at time 3, the buyer and the seller make simultaneous announcements of the state, and the outcome is determined according to rules 1 and 2 described in Section 2.7. In this case, the collusion possibilities at time 0 are non-trivial.

Theorem 2. *The game induced by the augmented secret message mechanism has a collusion-free perfect Bayesian equilibrium which produces the first-best outcome. Moreover,* all *perfect Bayesian equilibria produce the first best outcome.*

We prove this theorem on full implementation by proving two claims.

Claim 1. There exists a collusion-free perfect Bayesian equilibrium of the game induced by the augmented secret message mechanism where the outcome is first-best.

[17] Ties are broken arbitrarily, say in favor of B.

Proof of claim 1. The equilibrium strategies specify that both B and S say 0 at time -1. After both have said 0, the equilibrium strategies are isomorphic to those described in Section 2.7. Thus, by the arguments in that section, there exists a collusion free continuation equilibrium that produces the first best outcome.

If some player should say anything else than 0 at time -1, then they play a continuation equilibrium induced by the null mechanism. In this continuation equilibrium, B's expected payoff is $B_\emptyset$ plus whatever transfer received at time -1, and S's expected payoff is $S_\emptyset$ plus whatever transfer received at time -1. But at time -1, either S pays τ_B to B, or B pays τ_S to S. If S pays τ_B to B, then B's expected payoff is $B_\emptyset + \tau_B = B^*(t_B^*)$, from (10). Thus, B gets exactly $B^*(t_B^*)$, which is what he would get if both had said 0 (and S gets less than $S^*(t_S^*)$, because the total surplus will be less than first best). Similarly, if B pays τ_S to S then neither agent is better off than he would be if both had said 0. It follows that neither B nor S has any incentive to deviate and say anything else than 0 at time -1. This proves claim 1.

Claim 2. All perfect Bayesian equilibria produce the first best outcome.

Proof of claim 2. The buyer and seller can guarantee themselves the payoffs $B^*(t_B^*)$ and $S^*(t_S^*)$, respectively, by announcing a high integer at time -1. Therefore, in any perfect Bayesian equilibrium, the payoffs must be at least this high. They cannot be strictly greater, since the third party never pays. Thus, in all perfect Bayesian equilibria, the payoffs must be at the first-best level. This proves claim 2.

3. MORAL HAZARD IN TEAMS

A team consists of two agents, B and S. At time 1, B and S choose effort levels $b \geq 0$ and $s \geq 0$, respectively. Neither agent observes the other agent's effort. However, the team's total output $x \in \mathbf{R}$ is publicly observable. Output is a deterministic function of effort, $x = x(b, s)$. For simplicity there is no stochastic shock (it could easily be added). Assume $x(b, s)$ is increasing, concave and differentiable. The two main differences, compared to Section 2, is the unobservability of b and s, and the fact that the verifiable outcome x is a function of b and s (in Section 2, x was a real action implemented by the original judge).

Each agent is risk-neutral. B's cost of effort is $\varphi_B(b)$, and S's cost of effort is $\varphi_S(s)$. Each φ_i is increasing, differentiable and strictly convex, and $\varphi_i(0) = 0$. The first best action profile is

$$(b^*, s^*) \equiv \arg\max_{(b^*,s^*)\geq 0} \{x(b, s) - \varphi_B(b) - \varphi_S(s)\}.$$

The *first-best solution* specifies effort levels (b^*, s^*) and transfers t_B^* and t_S^*, such that $t_B^* + t_S^* = x(b^*, s^*)$. For the problem to be non-trivial, we assume $b^* > 0$ and $s^* > 0$. Individual rationality requires

$$B^* \equiv t_B^* - \varphi_B(b^*) \geq 0 \tag{12}$$

and

$$S^* \equiv t_S^* - \varphi_S(s^*) \geq 0. \tag{13}$$

If there is no message game, then as in Holmström's [12] pioneering article the original contract simply specifies transfers as a function of output x. Since ex post inefficient outcomes are renegotiated, it suffices to consider contracts that satisfy *budget balance*, that is, $t_B(x) + t_S(x) = x$ for all x. The budget-balance condition implies that it is impossible to implement the first-best without a third party (Holmström [12]). However, suppose there is a third party T who does not exert effort, does not observe any agent's effort, and whose transfer is $t_T(x)$. The budget-balance condition becomes $t_B(x) + t_S(x) + t_T(x) = x$. If there is no collusion then the first-best can be implemented by the following contract. For $i \in \{B, S\}$, let

$$t_i(x) = \begin{cases} t_i^* & \text{if} \quad x \geq x(b^*, s^*) \\ 0 & \text{if} \quad x < x(b^*, s^*) \end{cases} \tag{14}$$

The third party's transfer is $t_T(x) \equiv x - t_B(x) - t_S(x)$. This contract (weakly) implements (b^*, s^*) when collusion is not possible (Holmström [12]). However, collusion compromises this particular contract (Eswaran and Kotwal [6]). We now consider how collusion impacts other contracts, including those that ask for messages.

The time line is similar to the one described in Section 2.3. Thus, at time 0, there is a coalition-formation game where T can make a proposal to some player $i \in \{B, S\}$ to form coalition $C = \{i, T\}$. The collusive proposal specifies a set of "collusive mechanisms" $\{\Gamma_C(x)\}_{x \in \mathbf{R}}$. At time 1, agents B and S choose effort levels b and s, and joint output $x = x(b, s)$ is realized. Agent i's effort not observed by anyone except agent i, but the output x is publicly observed and is verifiable by outsiders. At time 2 nothing happens (there is no stochastic shock).

At time 3, the original mechanism Γ_0 is played out among the original parties. The mechanism specifies a message space M_i for each player $i \in \{B, S, T\}$. For each message profile $m \in M_B \times M_S \times M_T$ and output $x \in \mathbf{R}$, the transfers are $(t_B(m, x), t_S(m, x), t_T(m, x))$. We require $t_B(m, x) + t_S(m, x) + t_T(m, x) = x$ for all x.

At time 4, nothing happens if no coalition was formed at time 0. However, if a coalition C was formed, then the collusive mechanism $\Gamma_C(x)$ is played out among the colluding parties (where x is the output realized at stage 1). The collusive mechanism $\Gamma_C(x)$ specifies a message space $M_i^C(x)$ for each player $i \in C$. For each message profile $m^C \in \times_{i \in C} M_i^C(x)$ and output x, $\Gamma_C(x)$ specifies transfers $(t_i^C(m^C, x))_{i \in C}$. Here $t_i^C = t_i^C(m^C, x)$ is a monetary payment that agent $i \in C$ receives when messages m^C are sent at time 4 and x is the team output. Finally, at time 5 there is no scope for renegotiation, because the budget is balanced and there is no real decision to be made.

An effort profile (b, s) is (weakly) *implementable* if there is an original mechanism Γ_0 such that the induced game $G(\Gamma_0)$ has a PBE where the effort levels are (b, s). Since effort is unobserved, intuition suggests that the message game at time 3 is redundant. However, this intuition is incorrect. To see this, we first consider implementation without message games ($M_B = M_S = M_T = \emptyset$). We will show that in this case the third party plays no useful role. This generalizes Eswaran and Kotwal's [6] result in a way reminiscent of footnote 20 in Hart and Moore [9].

Theorem 3. *If we restrict attention to original contracts without messages, then introducing a third party does not expand the set of implementable effort profiles.*

To prove the theorem, suppose there is a third party T, and the effort profile $(\hat{b}, \hat{s})$ is implemented by Γ_0 without messages. We show that $(\hat{b}, \hat{s})$ can also be implemented without a third party. There are two cases, depending on whether or not collusion happens in equilibrium.

Case 1: The PBE of $G(\Gamma_0)$ which implements $(\hat{b}, \hat{s})$ is such that, in equilibrium, a side-contract is in force. To be specific, suppose coalition $C = \{B, T\}$ forms, with side-contract $\{\Gamma_C(x)\}_{x \in \mathbf{R}}$.

Notice that S will maximize his payoff only if, for all $s' \geq 0$,

$$t_S(x(\hat{b}, \hat{s})) - \varphi_S(\hat{s}) \geq t_S(x(\hat{b}, s')) - \varphi_S(s') \tag{15}$$

If B rejects T's proposal to collude, his payoff is sure to be

$$\mu \equiv \max_{b \geq 0} \{t_B(x(b, \hat{s})) - \varphi_B(b)\}$$

Indeed, S does not observe the collusive agreement, hence his choice of $\hat{s}$ is independent of it. Since B will accept any proposal that gives him more than μ, in fact B's equilibrium payoff must be exactly μ. The third party's equilibrium payoff must then be

$$t_B(x(\hat{b}, \hat{s})) + t_T(x(\hat{b}, \hat{s})) - \varphi_B(\hat{b}) - \mu \tag{16}$$

Claim 1. For all $b' \geq 0$,

$$t_B(x(\hat{b}, \hat{s})) + t_T(x(\hat{b}, \hat{s})) - \varphi_B(\hat{b}) \geq t_B(x(b', \hat{s})) + t_T(x(b', \hat{s})) - \varphi_B(b'). \tag{17}$$

Proof of claim 1. Suppose there is $b' \geq 0$ such that (17) is violated. Suppose T deviates from the equilibrium by offering B the following side-contract. If $x = x(b', \hat{s})$ then T pays B a side transfer $\hat{t}_B$ such that

$$t_B(x(b', \hat{s})) + \hat{t}_B - \varphi_B(b') = \mu + \varepsilon \tag{18}$$

where $\varepsilon > 0$. If $x \neq x(b', \hat{s})$ then B pays a big fine to T. (There are no messages in the side contract.) Since B only gets μ by rejecting, (18) implies that the unique sequentially rational response is to accept and choose b' so that the output is $x(b', \hat{s})$. Then T's payoff will be

$$t_T(x(b', \hat{s})) - \hat{t}_B = t_B(x(b', \hat{s})) + t_T(x(b', \hat{s})) - \varphi_B(b') - (\mu + \varepsilon) \tag{19}$$

where the equality uses (18). But, for small enough $\varepsilon > 0$, the violation of (17) implies that (19) is strictly greater than (16). Therefore, T is strictly better off by proposing the new side-contract, contradicting the definition of PBE. This proves the claim.

Suppose we get rid of the third party, and any transfer that T would have received is instead added to B's transfer, so B's transfer is $t_B(x) + t_T(x)$, for any x. Now (15) and (17) imply that this new mechanism implements $(\hat{b}, \hat{s})$, so the third party is useless.

Case 2: The PBE of $G(\Gamma_0)$ which implements $(\hat{b}, \hat{s})$ is such that, in equilibrium, no side-contract is in force.

In this case, the proof of claim 1 again goes through, so for any $b' \geq 0$, (17) must hold. But then just as in case 1, we can get rid of the third party. This completes the proof of Theorem 3.

Theorem 3 implies that, if the original contract does not include a message game, then the first-best is unattainable even if a third party is available. We now show that an original mechanism with a message game can implement the first-best, if a third party is available. This particular mechanism Γ_0 will be called the *whistle blowing mechanism*. In this mechanism, only T sends a message at time 3, with message space $M_T = \{\varnothing, \beta, \sigma\}$. Message $\varnothing$ is interpreted as "stay quiet", β is interpreted as "blow the whistle on agent B", and σ is interpreted as "blow the whistle on agent S". The message is observed by B and S, but not by outsiders.

Recall that (t_B^*, t_S^*) are transfers which satisfy (12) and (13). The outcome function is as follows.

Rule 1. If $x = x(b^*, s^*)$, pay $t_i = t_i^*$ to each $i \in \{B, S\}$, and set $t_T = 0$.

Rule 2. If $x \neq x(b^*, s^*)$ and T reports β, then B is paid $t_B = -2F$, T is paid $t_T = x + F$ and S is paid $t_S = F > 0$.

Rule 3. If $x \neq x(b^*, s^*)$ and T reports σ, then S is paid $t_S = -2F$, T is paid $t_T = x + F$ and B is paid $t_B = F$.

Rule 4. If $x \neq x(b^*, s^*)$ and T reports $\varnothing$, then pay $t_T = x$ to T, and set $t_B = t_S = 0$.

The key idea is that if the output is not first-best and the third party "blows the whistle" on someone, then that person is punished (by rule 2 or rule 3). If B and S

expect that the third party will blow the whistle on them if they try to collude with him, collusion will be deterred. They will only want to collude with the third party if the collusive contract can deter whistle blowing. Conversely, collusion is prevented if every collusive contract induces a continuation equilibrium with whistle blowing.

Theorem 4. *We can choose F so that the game $G(\Gamma_0)$ has a collusion-free perfect Bayesian equilibrium which produces the first-best outcome.*

To prove the theorem, a collusion-free perfect Bayesian equilibrium, producing the first best outcome, is constructed as follows. At time 0, T's strategy specifies that no collusive proposal is made. At time 1, any agent who has not signed a collusive contract sets effort at the first-best level. At time 3, T stays quiet if no collusive contract is in force.

Notice that the third party cannot affect the outcome by blowing the whistle on either agent as long as $x = x(b^*, s^*)$ (by Rule 1). At time 1, Rules 1 and 4 imply that both agents want to choose the first-best actions, anticipating that a deviation will lead to the entire output being given to the third party. From this it follows that, if there is no collusion at time 0, then the proposed strategies are sequentially rational from time 1 on. The outcome is first best by construction.

It remains to specify behavior after a time 0 deviation where T makes a collusive proposal. To support the equilibrium, such a deviation should not be profitable. To be specific, suppose T makes a proposal to B (the argument is similar for a proposal to S). We construct the strategies so that if B accepts and coalition $C = \{B, T\}$ forms, then either B or T will get no more than their equilibrium payoff. If T gets no more than zero, it is certainly not profitable for him not to propose the coalition. If B gets no more than B^* from joining the coalition, then we stipulate that his equilibrium strategy is to reject the proposal, and this behavior is certainly sequentially rational. Knowing that B will reject, it is again not profitable for T to make the proposal.

So suppose B accepts the proposal. The coalition $C = \{B, T\}$ forms and a collusive agreement $\{\Gamma_C(x)\}_{x \in x}$ is in force. S is unaware of a deviation, and will play as described above, i.e., his effort is s^*. The equilibrium strategies need to specify how the colluding players B and T will behave in $G(\Gamma_0)$ following the collusive deviation. Moreover, strategies should be such that either B gets less than B^*, or T gets less than 0. Also, we specify, if possible, that T and B believe that S chose the effort $s = s^*$ at time 1. In addition, if possible, T infers B's effort from the joint output, assuming $s = s^*$. In other cases, i.e., if x is inconsistent with $s = s^*$, then we may leave the beliefs unspecified.

For a given x, the collusive message game cannot be used to (uniquely) extract truthful information about whether or not T blew the whistle in Γ_0. This argument is the same as in Section 2.7. Since T's message in Γ_0 only changes the transfers,

the strategic incentives in the collusive mechanism $\Gamma_C(x)$ do not depend on it. That is, whistle-blowing does not induce any "preference reversal" at time 4. Hence, the collusive mechanism Γ_C cannot be designed to (uniquely) extract information about whether or not T blew the whistle at time 3. By assumption, the collusive mechanism must induce *some* continuation equilibrium (T is not allowed to propose a badly behaved collusive mechanism that causes an existence problem). By the payoff-irrelevance argument, we may assume the continuation equilibrium strategies are such that the messages sent at time 4 by B and T only depend on x, not on whether or not T blew the whistle at time 3.

At time 3, we specify that T blows the whistle on B. As $F > 0$, this is sequentially rational for T.

Under this specification, if the coalition $C = \{B, T\}$ forms, either Rule 1 or Rule 2 will apply at time 3. In either case, if F is large enough, S will not get less than his equilibrium payoff S^*. But then, at least one of the colluding agents is not made strictly better off. As argued above, this implies that T does not gain by making the proposal. This completes the proof of Theorem 4.

The key to the equilibrium construction is the third party's behavior at time 3. He stays quiet at time 3 if no collusive contract is in force. But if he is part of a collusive deal, then he blows the whistle on the other party to the collusion. In order to be assured of a profitable collusion, the colluding agents should design a collusive mechanism where whistle blowing is punished, and therefore not attractive to T, in all continuation equilibria. However, this is impossible because blowing the whistle only triggers a monetary reward which is not "payoff relevant" at time 4. Thus, all collusive mechanisms *must* have a continuation equilibrium where whistle blowing is not punished, and in such a continuation equilibrium the third party may as well blow the whistle. We support the collusion-free perfect Bayesian equilibrium of $G(\Gamma_0)$ by selecting the "whistle blowing continuation equilibrium" whenever a coalition is formed.

As before, full implementation is easily achieved as well.

REFERENCES

[1] Philippe Aghion, Mathias Dewatripont, and Patrick Rey (1994): "Renegotiation Design with Unverifiable Information," *Econometrica*, 62, 257-82.

[2] Sandro Brusco (1997): "Implementing Action Profiles when Agents Collude," *Journal of Economic Theory*, 73, 395-424.

[3] Yeon-Koo Che and Donald Hausch (1999): "Cooperative Investments and the Value of Contracting," *American Economic Review*, 89, 125-147.

[4] Yeon-Koo Che and József Sákovics (2004): "A Dynamic Theory of Hold-Up," *Econometrica*, 72, 1063-1104.

[5] Aaron Edlin and Stefan Reichelstein (1996): "Hold-Ups, Standard Breach Remedies and Optimal Investment," *American Economic Review*, 86, 478-501.

[6] Mukesh Eswaran and Ashok Kotwal (1984): "The Moral Hazard of Budget-Breaking," *RAND Journal of Economics*, 15, 578-581.

[7] Sanford Grossman and Oliver Hart (1986): "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 94, 691-719.

[8] Oliver Hart (1995): *Firms, Contracts and Financial Structure*, Oxford University Press: Oxford.

[9] Oliver Hart and John Moore (1988): "Incomplete Contracts and Renegotiation," *Econometrica*, 56, 755-785.

[10] Oliver Hart and John Moore (1990): "Property Rights and the Nature of the Firm," *Journal of Political Economy*, 98, 1119-58.

[11] Oliver Hart and John Moore (1999): "Foundations of Incomplete Contracts," *Review of Economic Studies*, 66, 115-138.

[12] Bengt Holmström (1982): "Moral Hazard in Teams," *Bell Journal of Economics*, 13, 74-91.

[13] Jean-Jacques Laffont and David Martimort (1997): "Collusion under Asymmetric Information," *Econometrica*, 65, 875-911.

[14] Eric Maskin and John Moore (1999): "Implementation with Renegotiation," *Review of Economic Studies*, 66, 39-56.

[15] Eric Maskin and Tomas Sjöström (2003): "Implementation Theory," *Handbook of Social Choice and Welfare*, Vol. 1, K. Arrow, A. Sen, and K. Suzumura (eds.), North-Holland, 237-288.

[16] Eric Maskin and Jean Tirole (1999): "Unforeseen Contingencies and Incomplete Contracts," *Review of Economic Studies*, 66, 83-114.

[17] Eric Maskin and Jean Tirole (1999): "Two Remarks on the Property Rights Literature," *Review of Economic Studies*, 66, 139-150.

[18] Georg Nöldeke and Klaus Schmidt (1995), "Option Contracts and Renegotiation: A Solution to the Hold-Up Problem," *Rand Journal of Economics*, 26(2):163-79..

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

[20] Ilya Segal and Michael Whinston (2002): "The Mirrlees Approach to Mechanism Design with Renegotiation (with Applications to Hold-Up and Risk-Sharing)," *Econometrica*, 70, 1-46.

[21] Tomas Sjöström (1996): "Implementation and Information in Teams", *Economic Design* 1: 327-341

[22] Jean Tirole (1986): "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations", *Journal of Law, Economics and Organization*, 2, 181-214.