

Implementing Action Profiles when Agents Collude

Sandro Brusco*

*Institut d'Anàlisi Econòmica, Campus Universidad Autònoma de Barcelona,
08193, Bellaterra, Barcelona, Spain*

Received September 15, 1994; revised May 17, 1996

In this paper we consider the principal–multiple agents problem when collusion among agents is possible. Collusion is captured by the use of equilibrium notions allowing for coalitional deviations. We first analyze the constraints that collusion puts on feasible wage schemes. Differently from the case of subgame perfect implementation the first best is not implementable. We provide a necessary condition that compensation schemes achieving the second best must satisfy. Roughly stated, the condition says that wage schemes that are collusion-proof are of the type “basic fixed wage plus collective bonus.” We next provide sufficient conditions for unique implementation. The conditions are analogous to those previously found for subgame perfect implementation, but an extra condition of Pareto-optimality (from the agents’ point of view) is needed. The mechanism we propose involves only a finite set of messages. No “integers game” or “tail chasing” construction is used. *Journal of Economic Literature* classification number: C7, D7, L2. © 1997 Academic Press

Press

1. INTRODUCTION

The general framework considered in this paper is the following. A principal has to instruct a group of agents to take a given action profile. The actual action performed by each agent is not observed by the principal, but it is observed by the colleagues. Each action profile induces a probability distribution on the revenue of the firm, and the realization of revenue is observable and verifiable. Furthermore, the principal can ask agents to send messages, and these messages are verifiable. Let a^* be a given action profile, and w^* a vector of compensation schemes (a compensation scheme is a function from the set of possible revenue realizations to the real numbers, i.e., payments can be made contingent on realized revenue). The

* The comments of a referee and an associate editor have greatly helped to improve both content and presentation of the paper. I remain responsible for all remaining errors. Financial support from the Spanish Ministry of Education, Proyecto de la DGICYT PB92-1138 is gratefully acknowledged.

implementation problem is: under what conditions can the principal design a game such that (a^*, w^*) is the unique equilibrium outcome?

It is obvious that the answer to this question depends on the particular equilibrium notion that we think appropriate for the model we are handling. Previous work has provided conditions for implementation when the equilibrium notion is Nash equilibrium or subgame perfect equilibrium (see Ma *et al.* [9] and Ma [8]). In this paper we want to explore conditions for implementation when the possibility of collusion among agents is taken into account.

The mechanisms presented in the cited papers are prone to manipulation by coalitions of agents. For example, suppose that a risk neutral principal wants risk averse agents to take costly efforts. The first best for the principal would be to pay agents a constant wage. It is shown by Ma [8] that, provided agents observe each other's effort, the principal can build a mechanism where the unique subgame perfect equilibrium outcome is that agents take the costly effort and receive the constant wage. If collusion is allowed, this outcome cannot survive. The reason is that any mechanism used for implementation must include the possibility for the agents to send a message (like "everybody took the prescribed action") leading to the desired compensation schedule, i.e. constant wages. The grand coalition of all agents can destroy the equilibrium in which the costly effort is taken by switching to a strategy profile in which all agents shirk and then the message "everybody took the prescribed action" is sent.

However, the exact meaning of "collusion" has to be carefully specified if we want to analyze the problem from the point of view of implementation theory. Different notions of collusion correspond to different solution concepts to be adopted in order to predict the outcome of a mechanism. The literature is rich with refinements of Nash equilibrium meant to capture the possibility of cooperation among agents. Recent examples are coalition-proof equilibrium (Bernheim *et al.* [3]), renegotiation-proof equilibrium (Farrell and Maskin [5]), far-sighted strong equilibrium (Li [7]). The narrowest, and also the oldest, is the concept of strong Nash equilibrium (see Aumann [2] and Maskin [10]).

The main difference between the various solution concepts is in the class of coalitional deviations which is allowed. The notion of Strong Nash equilibrium does not impose any constraint on the class of possible coalitional deviations. Thus, a strategy profile is a Strong Nash equilibrium if there is no coalition which, by adopting a different strategy profile, can increase the welfare of each member of the coalition. Other solution concepts restrict the class of feasible coalitional deviations using some "credibility" requirement (see Bernheim *et al.* [3] for a discussion).

In this paper we do not want to impose constraints on the possibility of collusion among agents, and we will therefore adopt the notion of strong

Nash equilibrium (or, more exactly, the appropriate refinement for extensive form games) as our solution concept. Furthermore, we allow agents to collude by agreeing on side transfers.

We are interested in implementation with *sequential* mechanisms, i.e. mechanisms where agents may be repeatedly called to send messages. This is useful only if we add some requirement of sequential rationality to the equilibrium notion we are adopting. In Section 2 we will define the appropriate refinement for extensive form games of the equilibrium notion under consideration. The reason why we consider sequential mechanisms is that in this case the results in Ma [8] provide a very useful benchmark. In that paper it is shown that the first best can be attained when the solution concept is subgame perfect equilibrium. The same result need not be true when Nash equilibrium is adopted. By adopting a sequential mechanism we want to find the departure from first best induced only by collusion, and not by the adoption of simultaneous vs. sequential mechanisms.

The paper provides two main results. The first one, contained in Section 3, is about the characterization of collusion-proof compensation schemes. As previously pointed out, it is clear that the first best (constant wages) is not implementable if a costly effort is required by some agent. We show that agents might be given a variable wage schedule which would however not be enough to induce them to take the required effort if no monitoring occurred. This rationalizes the urge of variable wages even in large organizations where individual effort has a negligible effect on the outcome. Variable wage schemes in such environments should not be seen as providing individual incentives (which are instead kept in line by monitoring) but rather “group incentives”, and their use appears to be appropriate when there is a strong possibility of collusion among agents.

An analogous result on compensation schemes in the presence of collusion has been obtained by Itoh [6]. There are two main differences between the approach taken here and the one in Itoh’s paper. First, Itoh assumes that the agents are able to sign binding contracts prescribing the level of effort to be chosen, while we only allow binding contracts based on verifiable variables; this includes firm’s revenue but excludes agents’ efforts. However, differently from Itoh, we allow the principal to build a revelation mechanism to find out the true effort profile. While this does not improve the principal’s welfare when agents can side contract on the level of effort (see Section 5 in Itoh’s paper), it becomes an effective way of enforcing a given action profile in our framework. In a way, reporting substitutes for binding contracts on effort, because an agent can force the other to take the prescribed level of effort by credibly threatening to report a deviation to the principal. The fact that the optimal compensation scheme is analogous under the two formulations shows that the two alternative ways to induce a given effort profile (i.e., binding contracts on effort signed by

the agents or reporting to the principal) are equally effective. The second difference between the two papers is that Itoh restricts attention to the case of two agents. When collusion is possible this is not an innocuous simplification, because it completely overlooks the threat posed by subcoalitions of agents. We analyze the general case of $I \geq 2$ agents, taking explicitly into account the possibility that subcoalitions are formed.

The second main result of the paper, contained in Section 4, is about implementation. We provide conditions on (a^*, w^*) that allow the principal to design a mechanism such that (a^*, w^*) is the unique equilibrium outcome. In particular we show that the condition of "preference reversal" used by Ma for subgame perfect implementation, together with Pareto efficiency (from the agents' point of view) is sufficient for implementation. By Pareto efficiency we mean that *when agents are paid according to the desired compensation schedules* there is no different action profile, together with a transfer scheme making each agent better off.

The addition of a Pareto-efficiency condition comes as no surprise, since it is known (see, e.g., Dutta and Sen [4]) that a social choice correspondence cannot be implemented in strong Nash equilibrium if it is not weakly Pareto-efficient. Clearly, no equilibrium can support a Pareto-inferior outcome because agents could jointly deviate and get the Pareto-superior outcome. Our result states that once Pareto-efficiency is added to the condition for subgame perfect implementation, then it is possible to design a mechanism implementing (a^*, w^*) . In other words, we can completely neutralize the threat coming from the possibility of collusion by a *subgroup* of agents, a result which was not obvious. Section 5 contains concluding remarks.

2. ON THE NOTION OF EQUILIBRIUM

Let Γ be an extensive form game (see Abreu and Sen [1] and Selten [13] for a more formal definition of extensive form games), and denote with Γ' a subgame of Γ . We denote by I the set of players, by S_i the set of strategies available to agent $i \in I$ and, for each subset $C \subset I$ we define $S_C = \prod_{i \in C} S_i$ and $S_{-C} = \prod_{i \notin C} S_i$.

If D is the set of decisions to be taken, we can define the outcome function for mechanism Γ as $g: S \rightarrow D$. For each proper subgame Γ' of Γ we can define for each agent the set of strategies S'_i defined by the subgame and the outcome function $g: S' \rightarrow D$. Each agent i has preferences over D , represented by a utility function $U_i: D \rightarrow \mathfrak{R}$.

Given any set X we denote by ΔX the set of probability measures on X . A mixed strategy for agent i , denoted σ_i , is just an element of ΔS_i . A mixed strategy for coalition C is an element of $\prod_{i \in C} \Delta S_i$, i.e. the Cartesian

product of mixed strategies of agents belonging to coalition C . In general, given a strategy profile $\sigma \in \prod_{i \in I} \Delta S_i$ we will denote with σ_C the marginal of σ with respect to coalition C . Thus, $\sigma_C(s_C)$ denotes the probability that coalition C takes action profile s_C .

Assuming that each U_i satisfies the expected utility hypothesis, we can represent preferences over (mixed) strategy profile σ by taking the appropriate expectation, i.e. we define $\bar{U}_i(\sigma) = E_\sigma[U_i(g(s))]$, where expectation on S is evaluated adopting the probability distribution induced by σ . Analogously, we define the utility for agent i of a deviation $\hat{\sigma}_C$ by coalition C as

$$\bar{U}_i(\hat{\sigma}_C, \sigma_{-C}) = \sum_{s_C \in S_C} \left(\sum_{s_{-C} \in S_{-C}} \sigma_{-C}(s_{-C}) U_i(g(s_C, s_{-C})) \right) \hat{\sigma}_C(s_C)$$

We can now define the equilibrium notions we are going to use.

DEFINITION 1. A strong Nash equilibrium (SNE) is a strategy profile $\sigma \in \prod_{i \in I} \Delta S_i$ such that for each coalition C and strategy $\hat{\sigma}_C \in \prod_{i \in C} \Delta S_i$ there exists at least one agent $i \in C$ such that $\bar{U}_i(\hat{\sigma}_C, \sigma_{-C}) \leq \bar{U}_i(\sigma)$.

Once we have defined the equilibrium concept, it is easy to add the requirement of subgame perfection for extensive games by imposing that the equilibrium strategy profile remains an equilibrium strategy profile in each proper subgame of the game.

DEFINITION 2. A Strong Perfect Equilibrium (SPE) is a strategy profile $\sigma \in \prod_{i \in I} \Delta S_i$ which is a Strong Nash Equilibrium for each subgame Γ' .

The notion of SPE has been first introduced by Rubinstein [12] for repeated games with infinite horizon, but it generalizes straightforwardly. We now proceed to the analysis.

3. OPTIMAL WAGES WHEN AGENTS COLLUDE

We consider a finite set I of agents and, when no confusion arises, we will take the liberty of denoting with the same symbol a set and its cardinality. Since we are looking at a multiple agents problem, it is assumed $I \geq 2$. Each agent can take an action $a_i \in A_i$, a finite set. Let $A = \prod_{i \in I} A_i$ be the set of all possible action profiles, with generic element a . The revenue of the firm is a random variable whose distribution depends on a and it is denoted by $\tilde{r}(\cdot | a)$. It will be assumed that the support of $\tilde{r}(\cdot | a)$ is finite and does not depend on a . Let us denote by R the set of possible realizations of revenue.

All agents have the same utility function

$$U(a, m) = V(m) - G(a_i),$$

where $V: \mathfrak{R} \rightarrow \mathfrak{R}$ is differentiable, strictly increasing, concave and unbounded below. We will assume that for each agent i there exists an action $\underline{a}_i \in A_i$ (“shirking”) such that $G(\underline{a}_i) < G(a_i)$ for each $a_i \in A_i \setminus \underline{a}_i$. Reservation utility is defined as $\underline{U} = V(0) - G(\underline{a}_i)$ and U is normalized to get $\underline{U} = 0$. The assumption that all agents have the same utility function is only made for convenience, and it can be easily relaxed. The assumption that V is unbounded below is also made for convenience, and could be replaced by weaker assumptions. Essentially, we want to ensure that agents not taking the prescribed action can be “punished enough”. The mechanism that we will propose in the next section does not rely on arbitrarily large punishments.

A compensation scheme for agent i is a function $w_i: R \rightarrow \mathfrak{R}$ which maps from the set of revenue realizations into real numbers. If the firm is a risk neutral profit maximizer, the first best is to choose an action profile a^* and a vector of fixed wages (m_1, \dots, m_I) solving

$$\max_{a \in A, m_1, \dots, m_I} \left(\sum_{r \in R} p(r|a) r \right) - \sum_{j=1}^I m_j \quad (1)$$

s.t.

$$V(m_i) - G(a_i) \geq 0 \quad \text{for each } i.$$

If for some agent we have $a_i^* \neq \underline{a}_i$ (which is the interesting case) then, as previously pointed out, the first best is not implementable when collusion is allowed. We will therefore try to characterize the second-best contract. Before doing that, we have to clarify what kind of transfers are allowed among agents.

3.1. Sidepayments

A sidepayment is an agreement among a group of agents to enact certain monetary transfers when some contingency occurs. It is therefore important, in order to define the set of admissible sidepayments, to specify on what contingencies sidepayments can or cannot be made dependent on, and whether or not they are observable. The framework we will adopt is the following:

- Transfers can only be made contingent on verifiable variables. This includes the revenue of the firm and messages sent inside the organization, but not the actions. Agents agree on transfer contracts before taking actions, and each agent observes the contracts signed by her colleagues.

- The firm can (and will) obtain that any employee be barred from signing contracts where transfers are conditional on messages sent inside the organization. The way this is obtained is by making the offer to join the firm conditional upon the agent not having signed any contract involving actions to be taken or messages to be sent inside the organization.
- Contracts are registered at courts and can be observed, possibly at a cost, by the principal.

This framework corresponds to a legal environment where the firm cannot prevent the members of the organization from subscribing to contracts making monetary transfers contingent on the revenue of the firm (notice that if it were possible to condition the participation to the organization on the absence of transfer contracts, then the possibility of sidepayments would not pose any problem to the firm). For example, the firm cannot prevent the agents from exchanging shares of the firm or other derivatives based on firm's income. However, the firm is allowed to forbid contracts contingent on actions inside the organization which are at odds with actions prescribed by the contract. To understand what is meant by this, suppose that the firm requires that agents report truthfully about the action profile they observe. Then, a sidepayment contract promising a monetary transfer to agent i if she announces an untruthful action profile would be considered illegal and therefore unenforceable, because such a contract would interfere with the obligations of the agent toward the firm. Legal norms requiring that the parties of a contract act "in good faith" are widespread. The prohibition of sidepayments contracts requiring the employee to "cheat" the principal seems therefore reasonable enough.

The last point refers to observability of sidepayment contracts. Here the idea is the following. If agents report that everybody took the prescribed action then the principal has no reason to suspect any wrongdoing, and the "regular" contractual wage w^* is therefore paid. In all other cases, wrongdoing is suspected and the principal and the agents go to court in order to decide what wage each agent should receive. Since courts can observe sidepayment contracts among agents, wages can be modified taking this into account. In fact, the mechanism that we will propose in Section 4 for implementation exactly relies on the idea of undoing transfer schemes among agents when an out-of-equilibrium message is observed.

The assumption that agents can observe the contracts signed by other agents is pretty strong. The reason why it is introduced is that the absence of observability of the contracts would radically change the nature of the implementation problem. To admit the possibility of unobservable sidepayments amounts to insert extra stages where moves (in this case, the stipulation of transfer contracts) are unobservable. This implies that

subgame-perfectness requirements are not binding any more, because no proper subgame exists. In order to deal with unobservable transfers we should therefore modify our solution concepts, introducing a notion of sequential rationality for imperfect information games when coalitional deviations are allowed. This line of research is indeed interesting but it would take us too far. In this paper we will limit ourselves to the analysis of the observability case. We conjecture however that, once the proper solution concepts for the incomplete information case are introduced, the mechanism that we propose can be easily adapted.

We can now proceed and define formally a sidepayment contract. As we said, transfers can be made contingent upon the revenue of the firm. A *transfer function for agent i* is thus a function $\tau_i: R \rightarrow \mathfrak{R}$ that for each realization of firm's revenue says how much the agent is supposed to pay or receive. A *transfer scheme* τ is just a collection of transfer functions $\tau = (\tau_1, \dots, \tau_I)$. A transfer scheme τ is said to be *admissible* if $\sum_{i=1}^I \tau_i(r) \leq 0$ for each realization r .

We will assume that transfer contracts can be stipulated by agents any time before actions are taken.

3.2. *Second Best*

We now try to characterize the second-best compensation scheme when collusion is allowed. For the moment, we will only consider the constraints imposed by the possibility of collusion among the entire set of agents, and disregard coalitions made up by proper subsets of agents. That the possibility of deviations by the grand coalition of all agents should be taken into account is pretty obvious. It is less obvious that it makes sense to disregard the possibility of deviations by subcoalitions. The rationale for this is that we can hope to stop deviations by subcoalitions through an appropriate monitoring mechanism. In other words, it may be possible to set up a mechanism where any agent can reveal to the principal that a group of agents is misbehaving, and incentives are designed in such a way that reporting wrongdoing by other agents is optimal. If this is the case, the principal will be able to punish deviating subcoalitions, so that deviations by subcoalitions can be prevented. In Section 4 we will discuss more precisely under what conditions it is possible to create a monitoring mechanism which stops all deviations by subcoalitions.

The case is different for deviations by the grand coalition. In this case agents may agree to take the wrong action profile *and* report that no wrongdoing occurred. Since all agents are part of the deviation, the principal cannot hope to elicit the truth. The only way to prevent deviations by the grand coalition is to design the compensation scheme in such a way that some agent is worse off whenever the wrong action profile is taken.

As a consequence, we define the second best as the solution of the program

$$\max_{a \in A, w_1, \dots, w_I} \sum_{r \in R} p(r|a) \left(r - \sum_{j=1}^I w_j(r) \right) \quad (2)$$

subject to the following constraints:

1. For every probability distribution $q \in \prod_{i \in I} \Delta A_i$ such that $\Pr(a) \neq 1$ and admissible transfer scheme $\tau = (\tau_1, \dots, \tau_I)$ it is possible to find an agent j such that:

$$E[V(w_j(r))|a] - G(a_i) \geq \sum_{\tilde{a} \in A} q(\tilde{a}) \{ E[V(w_j(r) + \tau_j(r))|\tilde{a}] - G(\tilde{a}_i) \}$$

2.

$$E[V(w_j(r))|a] - G(a_i) \geq 0 \quad \text{for each } i.$$

The first constraint says that the grand coalition of all workers cannot decide to adopt a different probability distribution q on the set of possible action profiles together with a feasible transfer scheme τ such that the utility of each worker is strictly superior to the utility of the solution proposed by the principal. The second set of constraints is the standard individual rationality condition.

It is clear that these constraints are to be satisfied for a compensation scheme to be collusion-proof. For the moment we don't consider the possibility of deviations by subcoalitions, under the assumption that monitoring is effective in preventing them (as we said, precise conditions under which this is true will be presented in Section 4). The question we pose in this section is: What do compensation schemes that are collusion proof look like? In other words, we would like to characterize the solution of program 2. We will start by analyzing a very simple model and then we will generalize our results.

3.3. A Simple $2 \times 2 \times 2$ Model

Consider a simplified model with only two agents. Each agent can either take action \underline{a} (low effort) or \bar{a} (high effort). Output can take only two values, $\bar{r} > \underline{r}$. Let $p(a_1, a_2)$ be the probability of outcome \bar{r} when the effort of the first agent is a_1 and the effort of the second agent is a_2 . We assume that the following string of inequalities holds:

$$1 > p(\bar{a}, \bar{a}) = p(\bar{a}, \underline{a}) = p(\underline{a}, \bar{a}) > p(\underline{a}, \underline{a}) > 0.$$

In other words, perfect monitoring is impossible (each outcome has positive probability under any effort profile) and the production function is

symmetric in the efforts of the two agents. Furthermore, once an agent takes the high effort there is no point in asking the other agent to take high effort.

It is pretty clear that if the principal wants to implement the action profile (\bar{a}, \bar{a}) then the compensation schemes have to be variable. It is less obvious that if some agent is required to take the low level of effort then her compensation has still to be variable. We will show that when the choice of the principal is to implement (\underline{a}, \bar{a}) then it is not optimal that the low effort agent is given a constant wage, while the high effort agent is given a variable wage. This is to be contrasted, respectively, with the case of no *ex post* monitoring and with the case of no collusion. Without *ex post* monitoring, i.e. without using the information acquired by agents in the production process, the principal would have to respect the individual incentive compatibility constraint of each single agent. We will show that by using *ex post* monitoring the principal can do strictly better even if collusion is allowed. On the other hand, given the fact that the principal is using *ex post* monitoring, it is known that in absence of collusion the first best can be implemented (Ma [8]). Thus, collusion causes the principal to do strictly worse.

The proof will be by contradiction. Suppose that it is optimal to pay a constant wage to agent 1 (i.e. a compensation scheme w_1 such that $w_1(\bar{r}) = w_1(\underline{r}) = m$). It is pretty obvious that agent 2 should be given a variable wage, since if both wages were constant agent 2 could shirk buying the silence of agent 1. Let $(w_2(\bar{r}) = m_1, w_2(\underline{r}) = m_2)$ be the compensation scheme for agent 2. Define $\bar{p} = p(\underline{a}, \bar{a})$, $\underline{p} = p(\underline{a}, \underline{a})$. Suppose that the above described compensation schemes are collusion-proof and allow the principal to implement the desired action profile. Let $\hat{U}_1 = V(m) - G(\underline{a})$ be the utility obtained by agent 1 and $\hat{U}_2 = \bar{p}V(m_1) + (1 - \bar{p})V(m_2) - G(\bar{a})$ be the utility obtained by agent 2.

We now show that it is possible to find a pair of compensation schemes which implements the same objective but is less costly to the principal, thus contradicting the fact that the pair $[(m, m), (m_1, m_2)]$ is optimal.

The alternative compensation package that we propose is the following:

- Agent 1 is paid $(m_1 - k_1, m_2 - k_2)$, where (m_1, m_2) is the compensation scheme originally intended for agent 2 and k_1, k_2 are real numbers that satisfy the relation:

$$\bar{p}V(m_1 - k_1) + (1 - \bar{p})V(m_2 - k_2) - G(\underline{a}) = \hat{U}_1. \quad (3)$$

- Agent 2 is paid $(m + v_1, m + v_2)$, where m is the (fixed) wage originally intended for agent 1 and (v_1, v_2) satisfy the relationship

$$\bar{p}V(m + v_1) + (1 - \bar{p})V(m + v_2) - G(\bar{a}) = \hat{U}_2. \quad (4)$$

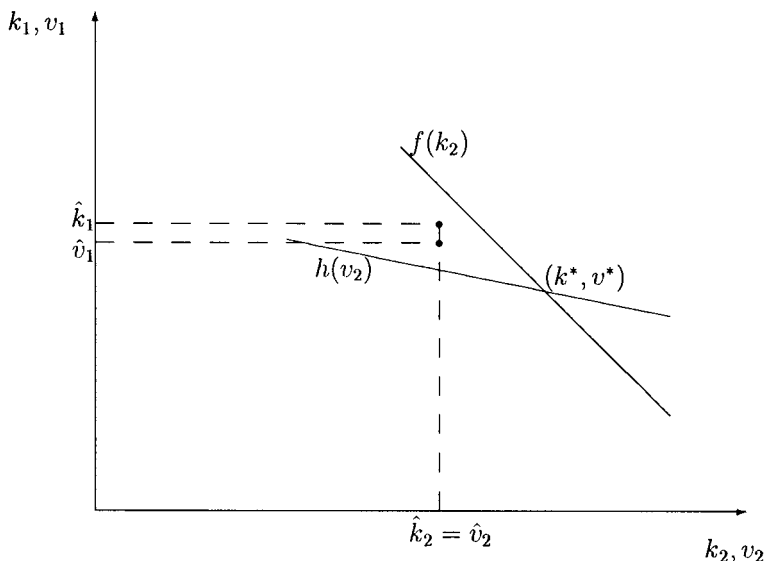


FIG. 1. The individual rationality constraint for agent 1 is satisfied below line f , and the individual rationality constraint for agent 2 is satisfied above line h . The pair (\hat{k}, \hat{v}) makes both agents strictly better off.

Thus, the idea is simply to switch the two compensation schemes between agents adjusting in order to keep both agents at the same utility level. The goal is now to show that we can find two pairs (k_1, k_2) and (v_1, v_2) such that the newly proposed compensation package is collusion proof and less costly to the principal. In particular, we will show that it is possible to find compensation schemes such that $k_1 > v_1$ and $k_2 = v_2$.

Let $k_1 = f(k_2)$ be the relation between k_1 and k_2 implicitly defined by (3), and analogously define $v_1 = h(v_2)$. Let us consider the behavior of f and h at the points $k_2^* = v_2^* = m_2 - m$. From (3) and (4) we have $f(k_2^*) = h(v_2^*) = m_1 - m$. In other words, at the points (k^*, v^*) the original contract is reinstated. We will now show that in a neighborhood of (k^*, v^*) we can find a pair (k, v) such that $k_1 > v_1$, $k_2 = v_2$ and both agents obtain at least as much utility as under the original compensation scheme. In order to do this, we observe that at k_2^* and v_2^* we have $f(k_2^*) = h(v_2^*)$. Thus, if we can show that $f'(k_2^*) \neq h'(v_2^*)$ we are done (see Fig. 1).

By differentiating (3) and (4) we obtain

$$\frac{dk_1}{dk_2} = -\frac{1 - \bar{p} V'(m_2 - k_2)}{\bar{p} V'(m_1 - k_1)}$$

and

$$\frac{dv_1}{dv_2} = -\frac{1 - \bar{p}}{\bar{p}} \frac{V'(m + v_2)}{V'(m + v_1)}.$$

Evaluating this derivatives at $k_2^* = v_2^* = m_2 - m$ and $k_1^* = v_1^* = m_1 - m$ we obtain

$$\left. \frac{dk_1}{dk_2} \right|_{k_1^*, k_2^*} = -\frac{1 - \bar{p}}{\bar{p}} \neq -\frac{1 - \bar{p}}{\bar{p}} \frac{V'(m_1)}{V'(m_2)} = \left. \frac{dv_1}{dv_2} \right|_{v_1^*, v_2^*}$$

The inequality follows from the assumption that $m_1 \neq m_2$ and the fact that V is strictly concave, so that $V'(m_1)/V'(m_2) \neq 1$. Therefore, it is possible to find (k, v) in a neighborhood of (k^*, v^*) such that $k_1 > v_1$ and $k_2 = v_2$ and each agent obtains at least \hat{U}_i .

We now show that any scheme of the type $(m_1 - k_1, m_2 - k_2), (m + v_1, m + v_2)$ such that $k_1 > v_1$ and $k_2 = v_2$ is collusion proof. Suppose not. Then it is possible to find a transfer (x_1, x_2) such that the two following inequalities hold:

$$\begin{aligned} \underline{p}V(m_1 - k_1 + x_1) + (1 - \underline{p})V(m_2 - k_2 + x_2) - G(\underline{a}) &> \bar{p}V(m_1 - k_1) \\ &+ (1 - \bar{p})V(m_2 - k_2) - G(\underline{a}) \end{aligned} \quad (5)$$

$$\begin{aligned} \underline{p}V(m + v_1 - x_1) + (1 - \underline{p})V(m + v_2 - x_2) - G(\underline{a}) &> \bar{p}V(m + v_1) \\ &+ (1 - \bar{p})V(m + v_2) - G(\bar{a}). \end{aligned} \quad (6)$$

Now recall that k and v are obtained in such a way that

$$\begin{aligned} \bar{p}V(m_1 - k_1) + (1 - \bar{p})V(m_2 - k_2) &= V(m) \\ \bar{p}V(m + v_1) + (1 - \bar{p})V(m + v_2) &= \bar{p}V(m_1) + (1 - \bar{p})V(m_2). \end{aligned}$$

Using the two equalities above and inequalities (5) and (6) we can show that the original compensation package was not collusion proof, a contradiction. In fact we have

$$\begin{aligned} \underline{p}V(m_1 - k_1 + x_1) + (1 - \underline{p})V(m_2 - k_2 + x_2) - G(\underline{a}) &> V(m) - G(\underline{a}) \\ \underline{p}V(m + v_1 - x_1) + (1 - \underline{p})V(m + v_2 - x_2) - G(\underline{a}) & \\ &> \bar{p}V(m_1) + (1 - \bar{p})V(m_2) - G(\bar{a}). \end{aligned}$$

If we can find a feasible transfer (y_1, y_2) transforming the original compensation scheme (i.e. $[(m, m), (m_1, m_2)]$) into the new compensation scheme

$$[(m_1 - k_1 + x_1, m_2 - k_2 + x_2), (m + v_1 - x_1, m + v_2 - x_2)],$$

we are done. Such a transfer is given by

$$y_1 = m_1 - m - v_1 + x_1, \quad y_2 = m_2 - m - v_2 + x_2.$$

Actually, this transfer makes agent 1 even better off than the original one, because v_1 is subtracted rather than k_1 .

3.4. A More General Case

The analysis of the simple case indicates what kind of constraints the possibility of collusive transfers imposes on the compensation scheme. We will now generalize the results obtained in the simple model.

Let $R = \{r_a, r_b, \dots, r_z\}$ be the set of possible revenue outcomes. For a given wage scheme w_i define $w_i(r_u) = m_u^i$ the amount paid to agent i when r_u occurs.

THEOREM 1. *Suppose that full monitoring is impossible, i.e. $p(r|a) > 0$ for each r and each action profile a , and suppose $a_i^* \neq \underline{a}_i$ for some i . Then the following conditions are necessary for optimality:*

1. $m_u^i \neq m_v^i$ for at least one pair of contingencies (r_u, r_v) .
2. For each pair of outcomes r_u and r_v and each pair of agents i, j we must have:

$$\frac{V'(m_u^i)}{V'(m_v^i)} = \frac{V'(m_u^j)}{V'(m_v^j)}.$$

Proof. The first point says that at least one agent has a variable wage.

This is obvious, given the assumption that $a_i^* \neq \underline{a}_i$ for some i . The second point implies that all wages are variable.

We now prove that point. Suppose the principal has chosen a vector of compensation schemes $w = (w_1, \dots, w_I)$ and an action profile a^* such that there exist two agents i and j and two contingencies r_u and r_v such that

$$\frac{V'(m_u^i)}{V'(m_v^i)} \neq \frac{V'(m_u^j)}{V'(m_v^j)}.$$

Since (w, a^*) is the solution of the maximization program, it must satisfy collusion proofness and individual rationality.

We will find a new vector of compensation schemes w' that is collusion proof, individually rational and implements a^* at a lower cost for the principal, thus contradicting the fact that (w, a^*) is a solution to the principal's problem.

Let \hat{U}_i, \hat{U}_j be the utilities attained by agent i and j in equilibrium. Consider the wage scheme \hat{w} defined as follows:

1. For agents other than i and j the compensation scheme is unchanged.

2. For agent i the compensation scheme is the same for contingencies other than u and v . For these contingencies we set $\hat{m}_u^i = m_u^i - k_u$ and $\hat{m}_v^i = m_v^i - k_v$, where (k_u, k_v) are set in such a way to satisfy:

$$E[V(\hat{w}_i)|a^*] - G(a_i^*) = \hat{U}_i.$$

3. For agent j the compensation scheme is the same for contingencies other than u and v . For these contingencies we set $\hat{m}_u^j = m_u^j + z_u$ and $\hat{m}_v^j = m_v^j + z_v$, where (z_u, z_v) are set in such a way to satisfy

$$E[V(\hat{w}_j)|a^*] - G(a_j^*) = \hat{U}_j.$$

If we can find two pairs (k_u, k_v) and (z_u, z_v) such that $k_u > z_u, k_v = z_v$ and the above described compensation scheme is collusion proof then we are done. The analysis is now identical to the $2 \times 2 \times 2$ case. ■

The necessary condition of Theorem 1 basically says that unless each agent is asked to "shirk" then at least one agent should have a variable wage, and if the wage of a single agent is variable, then the wage of any agent should be variable. This can be explained as follows. When collusive transfers are possible, it matters how the total wage bill (i.e., the sum of the wages of all employees) varies. This is because agents can agree to redistribute income among themselves to compensate the ones who can be damaged by an action profile different from the prescribed one. In order to avoid coalitional deviations, the total wage bill must shrink (in expectation) when the "wrong" action profile is taken. This makes sure that the increased utility that agents might have by performing less costly actions is offset by the decreased utility coming from a lower expected wage. This explains why the total wage bill must be variable.

The variability of the total wage bill works as a collective incentive compatibility constraint. The second point in Theorem 1 is an optimal insurance condition, and it is a consequence of the difference in risk attitudes of the principal and the agents. To see this, suppose we have just two agents and two contingencies. Let $(\bar{w}_i, \underline{w}_i)$ be the wage scheme of

player i and suppose that \bar{w}_i is paid with probability $p \in (0, 1)$. Consider the problem

$$\max_{\bar{t}, \underline{t}} pV(\bar{w}_1 + \bar{t}) + (1 - p) V(\underline{w}_1 + \underline{t})$$

$$\text{s.t. } pV(\bar{w}_2 - \bar{t}) + (1 - p) V(\underline{w}_2 - \underline{t}) = pV(\bar{w}_2) + (1 - p) V(\underline{w}_2).$$

The problem is, for given compensation schemes, to find a Pareto improving transfer while keeping the probability distribution unchanged. If we form the Lagrangian and write down the first-order conditions with respect to (\bar{t}, \underline{t}) we obtain

$$pV'(\bar{w}_1 + \bar{t}) - \lambda pV'(\bar{w}_2 - \bar{t}) = 0$$

$$(1 - p) V'(\underline{w}_1 + \underline{t}) - \lambda(1 - p) V'(\underline{w}_2 - \underline{t}) = 0.$$

Combining the two conditions we have

$$\frac{V'(\bar{w}_1 + \bar{t})}{V'(\underline{w}_1 + \underline{t})} = \frac{V'(\underline{w}_2 - \bar{t})}{V'(\underline{w}_2 - \underline{t})}.$$

For $\bar{t} = \underline{t} = 0$ to be a solution of the problem, the condition stated in Theorem 1 must be satisfied. Clearly, it is in the interest of the principal to design the compensation scheme in such a way that no Pareto improving transfers among agents are possible. Since the sum of transfers is zero for each contingency, the principal could incorporate them in the compensation scheme at no cost. Again, what matters is how the *aggregate* wage bill varies. Once the principal has made sure that variations of the total wage bill are enough to prevent collusive deviations, the way in which the aggregate risk is shared among agents follows optimal insurance criteria.

The content of Theorem 1 can be better understood if we take a look at the following corollary and remark.

COROLLARY 1.

• If $V(m) = -e^{-\alpha m}$ then the optimal wage scheme is a “fixed basic wage plus a fixed performance bonus equal for all.”

• If $V(m) = \log m$ or $V(m) = m^\alpha$, $\alpha \in (0, 1)$ then the optimal wage scheme is a “fixed wage plus a performance bonus proportional to the basic wage, with factor of proportionality equal for all.”

Proof. If $V(m) = -e^{-\alpha m}$ the form taken by the condition stated in Theorem 1 is

$$m_u^i - m_v^i = m_u^j - m_v^j$$

Thus, assuming $r_u > r_v$, we can find k such that $m_u^i = m_v^i + k$ and $m_u^j = m_v^j + k$. This means that the compensation scheme is made up by a “basic” (and possibly differentiated) wage m_0^i to be paid in any case, plus a fixed bonus k_u , identical for all agents, when revenue is $r_u > r_0$.

If $V(m) = \log m$ or $V(m) = m^\alpha$ the form taken by the condition is

$$\frac{m_u^i}{m_v^i} = \frac{m_u^j}{m_v^j}.$$

Thus, m_0^i is the “basic wage” of agent i , and if the output turns out to be $r_u > r_0$ then we can find β such that $m_u^i = \beta m_0^i$ for each i . ■

The particular form taken by the compensation scheme depends on the agents’ risk attitudes. With CARA utility functions, the absence of wealth effects makes sure that the optimal way to provide incentives is to give bonuses of the same absolute amount to each agent, while with power utility incentives have to be provided in the “proportional” form shown above.

REMARK. One interesting conclusion that we can draw from the above analysis is that in some cases compensation schemes that are not enough to provide individual incentives may be optimal after all because they prevent collusion. In fact, it is not difficult to build examples of collusion-proof compensation schemes which do not satisfy the incentive compatibility constraint of each single agent. By this we mean that if a single agent were to decide in isolation, without the fear that misbehavior might be reported to the principal, then the variability in her wage would not be enough to provide incentives to take the prescribed action.

It is sometimes observed that firms distribute bonuses related to overall performance, either fixed or proportional to the “basic wage.” Profit sharing schemes, for example, do exactly this. If the particular contribution of each worker to overall performance is next to negligible, then such schemes are puzzling, because they add variability to workers’ income without providing any incentive. The above analysis suggests that such schemes may be optimal after all, in that they provide incentives to agents not to collude. The “bonus” needed to prevent collusion may not be enough to provide individual incentives, yet the firm is better off.

4. UNIQUE IMPLEMENTATION

In the previous section we analyzed the constraints imposed by collusion on feasible compensation schemes. In this section we investigate whether

we can find a mechanism implementing a pair (a^*, w^*) satisfying the constraints.

There are two problems to be tackled. First, up to now we have ignored deviations by subcoalitions of agents and we only considered deviations by the grand coalition of all agents. The fact that the grand coalition of all agents cannot obtain an outcome which is Pareto-superior to (a^*, w^*) is clearly necessary, but not sufficient. The desired outcome must also be robust to deviations by subcoalitions of agents. Thus, the first task is to find a mechanism with the property that taking the action profile a^* is a strong perfect equilibrium when the compensation scheme is w^* .

Second, we want that the desired outcome be the *only* equilibrium outcome of our mechanism, that is we want to make sure that no other action profile can be undertaken by the agents in a SPE.

In this section we will provide conditions that allow us to build a mechanism where (a^*, w^*) is the unique SPE outcome. We will show that two conditions suffice. The first one is a condition of Pareto-efficiency. The second is the condition of "switching preferences" adopted by Ma for subgame perfect implementation (condition C.1 in Ma [8]). Since Pareto efficiency is an obvious requirement for implementation in solution concepts related to Strong Nash equilibrium, we find it interesting that implementation can be achieved by simply adding a condition which has also been adopted for subgame perfect implementation.

We first define formally the Pareto-efficiency condition. Let $\tilde{A} = \prod_{i \in I} \Delta A_i$ be the set of probability distributions on A that can be obtained when agents independently choose mixed strategies over A_i . For each $q \in \tilde{A}$ and transfer scheme τ define

$$U_i(q, \tau) = \sum_{a \in A} q(a) \{E[V(w_i^* + \tau_i) | a] - G(a_i)\}.$$

Thus, $U_i(q, \tau)$ is the expected utility of agent i when actions are taken according to the probability distribution q , the wage w_i^* is paid and the transfer scheme τ is agreed. In general, a pair (q, τ) denotes a probability distribution over A and a transfer scheme τ . We will denote by $(a^*, 0)$ the situation where a^* is taken with probability 1 and there are no transfers among agents. We assume that the following condition holds:

Condition 1 (Pareto Efficiency). For each pair $(q, \tau) \neq (a^*, 0)$ there is at least one agent i such that $U_i(q, \tau) < E[V(w_i^*) | a^*] - G(a_i^*)$

A condition of "weak Pareto efficiency" where the inequality in Condition 1 is weak is clearly necessary for implementation. If such a weak condition were not satisfied then the equilibrium yielding (a^*, w^*) would be vulnerable to the collective deviation where actions are taken as prescribed

by the probability distribution q , the message leading to w^* is adopted and the transfer scheme τ is agreed. Condition 1 strengthens this necessary condition. We have not been able to use the weak version in the proof of our implementation theorem. We feel however that there is little loss of generality in adopting the stronger version. We establish now two lemmas that will be useful in the proof of the main theorem of this section.

LEMMA 1. *If Condition 1 is satisfied then for each transfer scheme $\tau \neq 0$ it is possible to find $\delta(\tau) > 0$ with the property that for each probability distribution q there exists at least one agent i such that*

$$\sum_{a \in A} q(a) \{E[V(w_i^* + \tau_i + \delta(\tau)) | a] - G(a_i)\} < E[V(w_i^*) | a^*] - G(a_i^*).$$

Proof. Define the function

$$\gamma(q, \tau, \delta) = \min_{i \in I} \left\{ \left(\sum_{a \in A} q(a) \{E[V(w_i^* + \tau_i + \delta(\tau)) | a] - G(a_i)\} \right) - (E[V(w_i^*) | a^*] - G(a_i^*)) \right\}$$

Condition 1 implies that if $\tau \neq 0$ then $\gamma(q, \tau, 0) < 0$ for each $q \in \tilde{A}$. Define now the function $\hat{\gamma}(\tau, \delta)$ as

$$\hat{\gamma}(\tau, \delta) = \max_{q \in \tilde{A}} \gamma(q, \tau, \delta).$$

Notice that $\hat{\gamma}(\tau, \delta)$ is well defined, since \tilde{A} is a compact set and $\gamma(q, \tau, \delta)$ is continuous in q . Furthermore, we have $\hat{\gamma}(\tau, 0) < 0$. Since $\hat{\gamma}(\tau, \delta)$ is continuous in δ , it follows that we can find $\delta(\tau) > 0$ small enough such that $\hat{\gamma}(\tau, \delta(\tau)) < 0$. ■

Define $\bar{A} = \{q | q \in \tilde{A}, q(a^*) = 0\}$, that is \bar{A} is the set of probability distributions where at least one agent takes a_i^* with probability zero.

LEMMA 2. *If Condition 1 is satisfied and $\tau = 0$ then it is possible to find $\delta(0) > 0$ with the property that for each probability distribution $q \in \bar{A}$ there exists an agent i such that*

$$\sum_{a \in A} q(a) \{E[V(w_i^* + \delta(0)) | a] - G(a_i)\} < E[V(w_i^*) | a^*] - G(a_i^*)$$

Proof. The set \bar{A} is compact, and Condition 1 implies that for each q in this set we have $\gamma(q, 0, 0) < 0$. We can now apply the same reasoning as in Lemma 1. ■

The two lemmas are useful because they insure that it is possible to give limited rewards to an agent who “spies” and reveals to the principal that some agents deviated from the prescribed action. When collusion is possible this is important in order to avoid that the “right” equilibrium be destroyed by a group of agents who play a mixed strategy over actions and then receive with positive probability the reward. By keeping the reward low, this kind of manoeuvring is neutralized.

We now introduce a condition of “switching preferences” analogous to that introduced by Ma [8] (see Condition C.1 in his paper).

Condition 2 (Switching Preferences). Let \hat{a} and \tilde{a} be a pair of different action profiles. Then the probability distributions of $\tilde{r}(\cdot|\hat{a})$ and $\tilde{r}(\cdot|\tilde{a})$ are different, i.e. there exists $r \in R$ such that $p(r|\hat{a}) \neq p(r|\tilde{a})$.

The reason why Condition 2 is called “switching preferences” is that it implies that for each pair of action profiles (\tilde{a}, \hat{a}) it is possible to find two compensation schemes $w^y(\tilde{a}, \hat{a}), w^x(\tilde{a}, \hat{a})$ such that the two following inequalities hold:

$$\begin{aligned} E[V(w^y(\tilde{a}, \hat{a}))|\tilde{a}] &> E[V(w^x(\tilde{a}, \hat{a}))|\tilde{a}] \\ E[V(w^y(\tilde{a}, \hat{a}))|\hat{a}] &< E[V(w^x(\tilde{a}, \hat{a}))|\hat{a}]. \end{aligned}$$

In other words, when \tilde{a} is the true action profile then agents prefer to be paid according to $w^y(\tilde{a}, \hat{a})$ rather than according to $w^x(\tilde{a}, \hat{a})$. The opposite is true when \hat{a} is the true action profile. Therefore preferences between $w^y(\tilde{a}, \hat{a})$ and $w^x(\tilde{a}, \hat{a})$ switch when the true action profile switches from \tilde{a} to \hat{a} .

LEMMA 3. Consider a pair $w^y(\tilde{a}, \hat{a}), w^x(\tilde{a}, \hat{a})$ such that preferences switch. Let \hat{w} be a compensation scheme and define the two lotteries $\hat{w} \oplus w^y(\tilde{a}, \hat{a})$ and $\hat{w} \oplus w^x(\tilde{a}, \hat{a})$ as

with probability p payments are according to compensation scheme \hat{w} , and with probability $1 - p$ payments are according to $w^z(\tilde{a}, \hat{a})$, with $z = y, x$

Then for each \hat{w} and $p \in [0, 1)$ the two lotteries satisfy

$$\begin{aligned} E[V(\hat{w} \oplus w^y(\tilde{a}, \hat{a}))|\tilde{a}] &> E[V(\hat{w} \oplus w^x(\tilde{a}, \hat{a}))|\tilde{a}] \\ E[V(\hat{w} \oplus w^y(\tilde{a}, \hat{a}))|\hat{a}] &< E[V(\hat{w} \oplus w^x(\tilde{a}, \hat{a}))|\hat{a}]. \end{aligned}$$

Proof. It follows immediately from

$$E[V(\hat{w} \oplus w^y(\tilde{a}, \hat{a}))|\hat{a}] = pE[V(\hat{w})|\hat{a}] + (1 - p) E[V(w^y(\tilde{a}, \hat{a}))|\hat{a}].$$

Analogous equality holds when the compensation scheme is $w^x(\tilde{a}, \hat{a})$ and when expectation is conditional to \hat{a} . In fact, this is just the independence axiom for preferences over lotteries that can be represented by expected utility. ■

The reason why Lemma 3 is useful is that it allows us to reduce the set of collectively profitable deviations. The trick used in subgame perfect implementation to destroy possible equilibria where an undesired action profile \tilde{a} occurs together with a message profile leading to the adoption of w^* is to let any agent to denounce such an agreement, calling a test agent to choose w^y vs. w^x . If w^y is chosen then this is taken as evidence that \tilde{a} actually occurred, and the agent who spied is given a reward.

When subgame perfect implementation is adopted as a solution concept we don't need to worry about how w^y and w^* are ranked in the preferences of the test agent. But when coalitional deviations are taken into account these considerations become important. If w^y is preferred to w^* under a^* then two agents might destroy the "good" equilibrium by adopting the following coalitional deviation: Agent 1 claims that the true action profile is \tilde{a} , so that agent 2 becomes the test agent. Agent 2 in turn chooses w^y , which is preferred to w^* , so that agent 1 obtains the reward. Both agents would be strictly better off, and truth-telling would not occur. It is therefore important that compensation schemes used for "testing" give as low a utility as possible to the test agent. Lemma 3 assures us that this can be done without extra assumptions. In fact, in order to test that action profile \tilde{a} rather than \hat{a} has occurred, we can offer to an agent the choice between the lotteries $\hat{w} \oplus w^y(\tilde{a}, \hat{a})$ and $\hat{w} \oplus w^x(\tilde{a}, \hat{a})$. By taking \hat{w} with sufficiently low values and p (i.e., the probability that \hat{w} is paid) sufficiently close to one, we can make sure that the utility of these lotteries is less than the utility of w_i^* for each i , no matter what the true action profile is. For example, we can set $\hat{w}(r) = k$ for each r , where k is a real number defined by

$$k = \min_{i \in I, r \in R} w_i^*(r) - \varepsilon$$

with $\varepsilon > 0$. Since utility is strictly increasing, $V(k) < E[V(w_i^*)|a]$ for each a , and it is possible to pick p sufficiently close to one such that $E[V(k \oplus w^y)|a] < E[V(w_i^*)|a]$ for each a .

We now show that conditions 1 and 2 are sufficient for collusion-proof implementation.

THEOREM 2. *If conditions 1 and 2 are satisfied then (a^*, w^*) is implementable in SPE.*

Proof. We present here the mechanism used for the proof. In the Appendix it is shown that the only equilibrium outcome of this mechanism is the desired one. We first introduce some notation and definitions.

Let θ and β be two real numbers satisfying

$$V(\theta) - G(a_i) < V(\theta + \beta) - G(a_i^*) < 0$$

for each $a_i \in A_i$. The existence of θ and β is assured by the assumption that V is unbounded below. This is the only place where the assumption is used, and it is obvious that weaker assumptions would suffice.

For a given transfer scheme τ define τ_i^- as $\tau_i^-(r) = \min\{0, \tau_i(r)\}$ and τ_i^+ as $\tau_i^+(r) = \max\{0, \tau_i(r)\}$. Notice that $\tau_i = \tau_i^- + \tau_i^+$.

Concavity of V implies that

$$E[V(\theta + \beta + \tau_i^-(r)) - V(\theta - \tau_i^-(r)) | a] > G(a_i) - G(a_i^*)$$

for each action profile a and transfer scheme τ . To understand the condition, suppose that the principal is able to elicit the truth from the agents. Then the principal can punish agents who did not take the right action a_i^* with an heavy fine θ , and agents who take the proper action with a lighter fine $\theta + \beta$. Both fines are however nasty enough to push the agent below the reservation level of utility. The inequality remains true even when there exists a transfer scheme τ among the agents, provided that the principal can observe τ and subtract τ_i^+ from the compensation of each agent (remember that $\tau_i - \tau_i^+ = \tau_i^-$).

The condition makes sure that, *provided the principal can elicit the truth* each agent has an incentive to take the proper action no matter what the other agents are doing. Of course the proviso is important and it is the heart of the implementation problem.

Furthermore, for each transfer scheme τ and $\varepsilon > 0$ we can select $\hat{p}(\tau)$ close enough to one such that for each agent i

$$\hat{p}(\tau) V(\theta - \varepsilon(\tau) + \tau_i^-) + (1 - \hat{p}) E[V(w^z(\tilde{a}, \hat{a})) | \bar{a}] < V(\theta + \tau_i^-)$$

for each compensation scheme $w^z(\tilde{a}, \hat{a})$, $z = x, y$ and action profile \bar{a} . Again notice that $V(\theta) - V(\theta - \varepsilon) = \gamma$ implies $E[V(\theta + \tau_i^-) - V(\theta - \varepsilon + \tau_i^-)] \geq \gamma$ by concavity of V .

We can now describe the mechanism.

Stage 0. Any group of agents can decide to agree on a transfer scheme τ . These agreements are observed by each agent, but not by the principal, unless a cost c is paid.

Stage 1. Each agent chooses action $a_i \in A_i$. Action profile a is realized and observed by all agents. At this point the following mechanism is played.

Stage 2. Each agent announces an action profile, i.e. $m^i \in A$. The set of outcomes is partitioned as follows:

(a) $m^i = a^*$ for each i . In this case implement w^* .

(b) $m^i = \tilde{a} \neq a^*$ for each agent. In this case, the principal pays the cost c and learns the existing transfer scheme τ . Define

$$\tilde{\tau} = \min\{i \mid \exists \bar{z}: \forall j > \bar{z}, \tilde{a}_j \neq a_j^* \text{ and } \tilde{a}_k = a_k^* \text{ for each } \bar{z} \geq k \geq i\}.$$

Let $i(\tilde{a}) = \tilde{\tau}$ if it exists (i.e. at least one agent takes the prescribed action) and $i(\tilde{a}) = 1$ otherwise. Agent $i(\tilde{a})$ is paid $w_{i(\tilde{a})}^* + \delta(\tau)$. Other agents are paid $\theta + \beta - \tau_i^+$ if $\tilde{a}_i = a_i^*$ and $\theta - \tau_i^+$ otherwise. Notice that, since the transfer scheme τ has been agreed, the total payment for agent $i(\tilde{a})$ is $w_{i(\tilde{a})}^* + \tau_{i(\tilde{a})} + \delta(\tau)$, while for other agents is either $\theta + \beta + \tau_i^-$ or $\theta + \tau_i^-$.

(c) In all other cases, go to Stage 3.

Stage 3. All agents are asked to choose between $w^y(\tilde{a}, \hat{a})$ and $w^x(\tilde{a}, \hat{a})$ for each possible pair (\tilde{a}, \hat{a}) . In case there is an action \tilde{a} such that agent i always selects $w_i^y(\tilde{a}, \hat{a})$ over $w_i^x(\tilde{a}, \hat{a})$ for each \hat{a} (there can be at most one) we say that “agent i indicates \tilde{a} ”. Payoffs are as follows:

(a) If all agents indicate \tilde{a} then agent $i(\tilde{a})$ is paid $w_i^* + \delta(\tau)/2$. A pair (\tilde{a}, \hat{a}) is randomly extracted with identical probability on $A \times A$, and each agent is paid $\theta - \varepsilon - \tau_i^+$ with probability $\hat{p}(\tau)$ and the choice between $w^y(\hat{a}, \tilde{a}) - \tau_i$ and $w^x(\hat{a}, \tilde{a}) - \tau_i$ with probability $1 - \hat{p}(\tau)$.

(b) If all agents indicate a^* then we set $i(a^*) = 1$ if agent 1 sent message $m^i = a^*$ at Stage 2, and $i(a^*) = 2$ otherwise. Agent $i(a^*)$ is paid $w_i^* + \delta(\tau)/2$. A pair (\tilde{a}, \hat{a}) is randomly extracted with identical probability on $A \times A$, and each agent is paid $\theta - \varepsilon - \tau_i^+$ with probability $\hat{p}(\tau)$ and the choice between $w^y(\hat{a}, \tilde{a}) - \tau_i$ and $w^x(\hat{a}, \tilde{a}) - \tau_i$ with probability $1 - \hat{p}(\tau)$.

(c) In all other cases, a pair (\tilde{a}, \hat{a}) is randomly extracted with identical probability on $A \times A$, and each agent is paid $\theta - \varepsilon - \tau_i^+$ with probability $\hat{p}(\tau)$ and the choice between $w^y(\hat{a}, \tilde{a}) - \tau_i$ and $w^x(\hat{a}, \tilde{a}) - \tau_i$ with probability $1 - \hat{p}(\tau)$.

Notice that, since the transfer scheme τ has been agreed, the total payment for agent $i(\tilde{a})$ is $w_{i(\tilde{a})}^* + \tau_{i(\tilde{a})} + \delta(\tau)/2$, while for other agents is the lottery $(\theta - \varepsilon + \tau_i^-) \oplus w^z(\tilde{a}, \hat{a})$. ■

While the mechanism might at first sight look quite complex, its structure is actually very simple. Basically, agents are required to report the

action profile. If they all agree that the true action is a^* then w^* is paid, while if they all agree that $\tilde{a} \neq a^*$ is the true action profile then all agents but one are fined (we will discuss later why one agent should not be fined). In case agents report different action profiles, the principal tries to find the true action profile by offering the choice between different compensation schemes (this is what is going on in Stage 3). If it is still not possible to determine the true action profile then all agents are fined, otherwise one agent is given a prize (the same agent who would have obtained the prize under truth-telling) and all other agents are fined.

The reason why the principal does not want to fine *all* agents is that we have to make sure that truth-telling occurs when the true action profile is $\tilde{a} \neq a^*$. If we were to fine all agents when such a message is sent, then the coalition of all agents would be better off switching to the message $m^i = a^*$ for each i . In order to prevent this, we need that when \tilde{a} is the true action profile, at least one agent be better off when the truth is told. In our mechanism, this is agent $i(\tilde{a})$.

The reason way $i(\tilde{a})$ is established in such a complicated way, is to make sure that no subcoalition of at most $I-1$ agents can choose a deviation from a^* such that each member of the coalition received the price with positive probability. It can in fact be checked that for every proper subset $C \subset I$ it is true that when each agent in the complementary coalition $I \setminus C$ takes the prescribed action then at least one agent in C can never be $i(\tilde{a})$.

The reason why a “stage” game is used, with all agents sending messages simultaneously at each stage, is that in this way we increase the size of a coalition needed to perform a profitable deviation. To see this, consider the case $I=3$ and suppose that the mechanism is modified in the following way: At Stage 2, agent 1 announces $m^1 \in A$ first, followed by agent 2 and then by agent 3. The important modification is that now agent i knows the messages of agents $j < i$ when choosing the message. What can go wrong in this case? Consider the following deviation by the coalition made up by the first two agents: Agent 1 chooses a_1 rather than a_1^* and bribes agent 2 to report a^* . If agents 1 and 2 report a^* and the true action profile is (a_1, a_2^*, a_3^*) then the optimal thing to do for agent 3 is to announce a^* as well. Any other message would cause Stage 3 to be reached and, under truth-telling, agent 3 would be worse off.

While the problems created by this specific example could be taken care of by appropriately modifying other parts of the mechanism, the general message should be clear. By letting agents announce messages sequentially (as in Ma [8]) the possibility of profitable deviations by subcoalitions of agents is increased. A subcoalition may break the good equilibrium, leaving the remaining agents no choice but to go along with the lie.

This does not happen if messages are issued simultaneously, so that agents move at “stages”. On the other hand, simultaneity may create

coordination problems, but these are dealt with through the collusive solution notion. In case of uncoordinated messages, the mechanism delivers an outcome which is strictly Pareto inferior (meaning that there is an alternative message such that each agent is strictly better off). This makes sure that uncoordinated messages can never be part of an equilibrium. A connected observation is that the mechanism adopted in the proof does not guarantee subgame perfect implementation. There are subgame perfect equilibria where agents do not coordinate their messages and therefore the wrong outcome occurs. While these outcomes are Pareto inefficient, they cannot be broken by unilateral deviations. If we were willing to allow for "tail chasing" devices such as integer games or modulo games, the mechanism proposed in the proof of the theorem could be augmented in order to ensure implementation in subgame perfect equilibrium as well.

5. CONCLUSION

In this paper we have analyzed the principal-multiple agents problem when collusion is allowed. The interplay of collusion and monitoring produces interesting results with respect to the second best wage scheme. We have shown that it may be optimal for the principal to introduce variability in the compensation schemes of risk-averse agents even if the output is not particularly reactive to the action taken by a single agent. Individual incentives are kept in line by monitoring, not by non-constant wages. The role of variability in wages is to provide the agents *as a group* with the incentive to take the desired action profile. Individual incentives can be ignored because agents observe the whole action profile, and it is possible to design a mechanism such that any deviation from the prescribed action by a group of agents will be reported to the principal.

In this paper we have not analyzed the choice of the production process, and we have simply assumed that each agent could observe the entire action profile. Our analysis suggests that the way in which the production process is designed, and in particular the possibility of monitoring by other agents, has important consequences for the optimal compensation schemes. This in turn has implications for the optimal choice of the production process. Suppose, for example, that a principal can choose between two production processes, one in which each agent works in isolation and another in which actions are observable. For example, in order to make observable the effort of each agent the principal might institute frequent job rotation, organize offices as open spaces and so on. What organizational mode is optimal will depend not only on technological factors, but also on the induced differences in compensation scheme. If agents work in isolation the individual incentive compatibility constraints will have to be satisfied,

while in the second case only group incentives have to be provided. The study of the relationship between choice of the production process and choice of the compensation scheme appears to be an interesting topic for future research.

APPENDIX

Proof of Theorem 2

In this Appendix we show that, in the mechanism proposed, (a^*, w^*) is the only outcome supported by a strong perfect equilibrium. Lemma 4 shows that there exists a SPE supporting (a^*, w^*) , and Lemma 5 shows uniqueness.

LEMMA 4. *There exists a strong perfect equilibrium with outcome (a^*, w^*) .*

Proof. The following strategy profile has the required properties:

- Stage 0: No transfer contract is signed.
- Stage 1: Each agent takes the prescribed action a_i^* , no matter what transfer contract has been signed.
- Stage 2: For each action profile, each agent tells the truth.
- Stage 3: Each agent chooses the preferred wage schedule for each pair of action profiles (\tilde{a}, \hat{a}) .

We will show that the proposed strategy profile is a strong Nash equilibrium in each subgame.

Stage 3. In the equilibrium agent $i(\tilde{a})$ is already obtaining the best possible outcome of the subgame, and she can't be part of any deviating coalition. Coalitions of at most $I-1$ agents can only put the outcome in class 3(c). Since some deviators must be choosing not preferred wage schedules for the deviation to be effective, these agents are strictly worse off.

Stage 2. Suppose that a^* is the true action profile and that the transfer scheme τ has been agreed, and denote with E^* expectation conditional to a^* being the true action profile. We will first show that it is impossible to have a deviation involving all agents, i.e. at least one agent is worse off. Let p be any probability distribution on messages resulting from the deviation. First observe that agent I obtains no more than $E^*V(\theta + \beta + \tau_I^-)$ unless the message is in class (a) or the message is in classes (b) or (c) and $i(\tilde{a}) = I$. Let p^* be the probability that the message is in class (a) and p^I be

the probability that the message is in (b) or (c) and agent I is $i(\tilde{a})$. For the deviation to be viable the following inequality must hold:

$$p^I(E^*V(w_I^* + \tau_I + \delta(\tau))) + p^*(E^*V(w_I^* + \tau_I)) \\ + (1 - p^I - p^*)(E^*V(\theta + \beta + \tau_I^-)) > E^*V(w_I^* + \tau_I).$$

This in turn implies that

$$p^I > (1 - p^*) \frac{E^*V(w_I^* + \tau_I) - E^*V(\theta + \beta + \tau_I^-)}{E^*V(w_I^* + \tau_I + \delta(\tau)) - E^*V(\theta + \beta + \tau_I^-)}. \quad (7)$$

Consider now agent $I-1$. When agent I is $i(\tilde{a})$, agent $I-1$ obtains at most $E^*V(\theta + \beta + \tau_{I-1}^-)$. If the message is in class (a) then $E^*V(w_{I-1}^* + \tau_{I-1})$ is obtained and in all other cases at most $E^*V(w_{I-1}^* + \tau_{I-1} + \delta(\tau))$ is obtained. Thus, the following inequality must be satisfied

$$p^I(E^*V(\theta + \beta + \tau_{I-1}^-)) + p^*(E^*V(w_{I-1}^* + \tau_{I-1})) \\ + (1 - p^I - p^*)(E^*V(w_{I-1}^* + \tau_{I-1} + \delta(\tau))) > E^*V(w_{I-1}^* + \tau_{I-1}),$$

which in turn implies that

$$p^I < (1 - p^*) \frac{E^*V(w_{I-1}^* + \tau_{I-1} + \delta(\tau)) - E^*V(w_{I-1}^* + \tau_{I-1})}{E^*V(w_{I-1}^* + \tau_{I-1} + \delta(\tau)) - E^*V(\theta + \beta + \tau_{I-1}^-)}. \quad (8)$$

It is immediate to see that it is possible to choose $\delta(\tau) > 0$ small enough so that inequalities 7 and 8 cannot be simultaneously satisfied. We have therefore established that when a^* occurs no deviation from truth-telling involving all agents is possible.

Consider now deviations by a proper subset of agents. Since at least one agent is telling the truth, the message can never fall in class (b), so the only possibility is that Stage 3 is reached. At Stage 3 the non-deviating agents will indicate a^* , so that the only agent who can benefit from the deviation is $i(a^*)$, i.e. either agent 1 or 2. No deviation by a single agent is therefore profitable, since if agent 1 was the only deviator he would have to choose $m^1 \neq a^*$ in order to deviate. Consider a joint deviation by 1 and 2. The agent who does worse get at most $\frac{1}{2}E^*V(w_i^* + \tau_i + \delta(\tau)) + \frac{1}{2}E^*V(\theta + \tau_i^-)$, which is strictly less than $E^*V(w_i^* + \tau_i)$ for $\delta(\tau)$ small enough. Thus, the deviation is not profitable. We conclude that no deviation from truth-telling is possible when a^* has occurred.

Suppose next that $\tilde{a} \neq a^*$ is the true action profile. Agent $i(\tilde{a})$ is already obtaining the best possible outcome of the subgame and she will not deviate. Coalitions of at most $I-1$ agents can only move the outcome to

class (c), so that Stage 3 is reached. Since $i(\tilde{a})$ will tell the truth, the best outcome for the deviating coalition is that everybody indicates \tilde{a} , making all deviating agents worse off.

Stage 1. We first show that there is no deviation involving all agents. Consider first the case $\tau \neq 0$. In this case Lemma 1 implies that it is possible to find $\delta(\tau)$ such that for each $q \in \tilde{\mathcal{A}}$ we have

$$\min_{i \in I} \left\{ \sum_{a \in \mathcal{A}} q(a) U^i(a, w_i^* + \tau_i + \delta(\tau)) - U^i(a^*, w_i^*) \right\} < 0,$$

where $U^i(a^*, w_i^*) = E[V(w_i^*) | a^*] - G(a_i^*)$ and

$$U^i(a, w_i^* + \tau_i + \delta(\tau)) = E[V(w_i^* + \tau_i + \delta(\tau)) | a] - G(a_i).$$

This in turn implies that at least one agent is strictly worse off when a side payment $\tau \neq 0$ is agreed.

Suppose now that $\tau = 0$. Since we have to compare the outcome from the deviation to the equilibrium outcome, we can limit ourselves to consider deviations where taking a^* and having the message at Stage 2 in class (a) has probability 0. For a given deviation, let $q(a)$ be the probability that action profile a occurs. Let $U^i(a)$ be the expected payoff of agent i when action profile a occurs and define

$$\hat{U}^i = \sum_{a \in \mathcal{A} \setminus a^*} \frac{p(a)}{1 - p(a^*)} U^i(a).$$

By Lemma 2 it is possible to find $\delta(0) > 0$ such that for each $q \in \tilde{\mathcal{A}}$ it is possible to find agent j such that $\hat{U}^j < U^j(a^*, w_j^*)$. Define p^j as the probability that a^* occurs and agent j obtains $w_j^* + \delta(0)$ and notice that in these cases agent $j-1$ (or agent 2, if $j=1$) obtains less than reservation utility. Let $\tilde{U}^j = \max\{\hat{U}^j, \underline{U}^j\}$. For the deviation to be profitable the following inequality must hold,

$$p^j(U^j(a^*, w_j^* + \delta(0))) + (p(a^*) - p^j) \underline{U}^j + (1 - p(a^*)) \hat{U}^j > U^j(a^*, w_j^*),$$

which in turn implies that

$$p^j(U^j(a^*, w_j^* + \delta(0))) + (1 - p^j) \tilde{U}^j > U^j(a^*, w_j^*)$$

or

$$p^j > \frac{U^j(a^*, w_j^*) - \tilde{U}^j}{U^j(a^*, w_j^* + \delta(0)) - \tilde{U}^j}. \quad (9)$$

Consider now agent q , with $q = j - 1$ if $j > 1$ and $q = 2$ otherwise. Define

$$\bar{U}^q = \max\{\hat{U}^q, U^q(a^*, w_q^* + \delta(0))\}$$

and

$$\underline{U}^q = V(\theta + \beta) - G(\underline{a}_q).$$

Since agent q obtains less than \underline{U}^q when agent j receives $w_j + \delta(0)$, the following inequality must be satisfied for the deviation to be profitable to agent q

$$p^j \underline{U}^q + (1 - p^j) \bar{U}^q > U^q(a^*, w_q^*)$$

which in turn implies that

$$p^j < \frac{\bar{U}^q - U^q(a^*, w_q^*)}{\bar{U}^q - \underline{U}^q}. \quad (10)$$

Since $U^q(a^*, w_q^*) > \underline{U}^q$, for δ small enough inequalities (9) and (10) cannot hold together. We conclude that no deviation involving all agents is possible.

Consider now deviations involving a proper subset of agents. It is immediate to see that no unilateral deviation can be profitable, so let us consider coalitions of at least 2 agents. First observe that no deviation can involve taking action profile a^* at Stage 1 and then a message that does not fall in class (a), because the previous analysis applies (i.e., if a^* is the true action profile and at least one agent is telling the truth then the best that any coalition can do is to tell the truth).

Suppose now that a deviating coalition causes an action profile \tilde{a} to occur. Since the deviating coalition involves at most $I - 1$ agents, at least one agent must be telling the truth at Stage 2. Thus at Stage 2 the message cannot be in class (a), so we are left with classes (b) and (c). Since one agent is not deviating, truth-telling must occur when Stage 3 is reached. Thus, agent i can profit from the deviation only if $i = i(\tilde{a})$ with positive probability, because in all other cases utility is less than $U^i(a^*, w^*)$. Therefore, a deviation can be profitable for agent $i > 1$ only if agent $i - 1$ takes $\tilde{a}_i \neq a_i^*$ with positive probability (if not, agent i can never be $i(\tilde{a})$). This implies that whenever agent i belongs to the deviating coalition then each agent $j < i$ must belong to the deviating coalition, and in particular agent 1 must belong to the deviating coalition. Since we are considering deviations by coalitions of at least two members, agent 1 must always belong to the deviating coalition. Furthermore, it can't be the case that agent I belongs to the deviating coalition, since it has been assumed that it has at most $I - 1$ members, but if I does not belong to the coalition then she takes a_i^*

with probability 1. This implies that agent 1 can never be agent $i(\tilde{a})$, so that she cannot belong to the deviating coalition. This establishes a contradiction, so that no deviating coalition exists.

Stage 0. Since the action profile a^* and truth-telling will occur for any give transfer scheme τ then, it follows directly from condition 1 that any transfer scheme $\tau \neq 0$ will make at least one agent worse off. ■

LEMMA 5. *The only outcome supported by a SPE is (a^*, w^*) .*

Proof.

Stage 3. Let \tilde{a} be the true action profile, possibly with $\tilde{a} = a^*$. Consider any probability distribution on messages at Stage 3. It must be the case that agent $i(\tilde{a})$ chooses wage schedules indicating an action profile \bar{a} such that $i(\bar{a}) = i(\tilde{a})$. If not, she would be strictly better off by choosing the preferred wage schedules for each pair, thus indicating \tilde{a} . This in turn implies that whenever an action is unanimously indicated by all agents, agent $i(\tilde{a})$ is selected. But in this case each agent $i \neq i(\tilde{a})$ is strictly better off indicating the true action profile with probability one. Thus, the only equilibrium outcome at Stage 3 is that each agent indicates the true action profile.

Stage 2. Suppose $\tilde{a} \neq a^*$ is the true action profile. The outcome can never fall in class (a), because agent $i(\tilde{a})$ would be strictly better off by replacing the message $m^i = a^*$ with $m^i = \tilde{a}$. In this case Stage 3 would be reached and truth-telling would occur, making $i(\tilde{a})$ strictly better off. Therefore, only (b) or (c) are possible equilibrium outcomes. Furthermore, it can never be the case that outcome (b) is reached under a message \bar{a} such that $i(\bar{a}) \neq i(\tilde{a})$. In that case, agent $i(\tilde{a})$ would be strictly better off by replacing \bar{a} with \tilde{a} . We can therefore conclude that the only possible equilibrium outcomes are (b) with a message \bar{a} such that $i(\bar{a}) = i(\tilde{a})$ or (c) with truth-telling at Stage 3. The latter case is Pareto-inferior to the former. Therefore, the only equilibrium outcome is that the message is in (b) and \tilde{a} obtains $w_{i(\tilde{a})}^* + \tau_i + \delta(\tau)$.

Assume now that the true action profile is a^* . It can never be the case that in equilibrium the message is in class (b) with a message \bar{a} such that $i(\bar{a}) \neq 1$. In this case agent 1 would be strictly better off by replacing \bar{a} with a^* . This would move the outcome from (b) to (c), allowing 1 to get $w_1^* + \delta(\tau)/2 + \tau_1$ rather than at most $\theta + \beta + \tau_1^-$, or from (c) with 1 obtaining $\theta - \varepsilon + \tau_1^-$ to either (a) or (c) with 1 obtaining $w_1^* + \delta(\tau)/2 + \tau_1$. We conclude that in equilibrium agent 1 announces either a^* or \bar{a} with $i(\bar{a}) = 1$ whenever the true action profile is a^* , and in equilibrium it is never the case that the outcome is in class (b) with a message \tilde{a} with $i(\tilde{a}) \neq 1$. However, it is also impossible that the message is in class (b) with $i(\tilde{a}) = 1$,

because in this case agent 2 would be strictly better off replacing \tilde{a} with a^* . This would move the outcome from (b) to (c) with $i(a^*) = 2$ (because this only occurs if 1 is reporting something different from a^*). If the original outcome is in (c) then agent 2 is strictly better off if the outcome moves to (a), and not worse off if the outcome remains in (c). Therefore, outcome (b) with $i(\tilde{a}) = 1$ can never occur in equilibrium.

Suppose now that with positive probability the outcome is in class (c). In this case the message of agent 1 must be a^* with probability 1. But this in turn implies that the coalition of the remaining $I - 1$ agents is strictly better off switching to $m^i = a^*$ with probability 1 as well. Thus (a) is the only possible outcome.

Stage 1. We have previously shown that truth-telling occurs at Stage 2 for each action profile. Since truth-telling occurs at Stage 2, each agent $i > 1$ is strictly better off taking action a_i^* rather than any other action a_i . Since all agents other than 1 take a_i^* , it must be the case that agent 1 takes a_1^* as well.

Stage 0. Since the action profile a^* is going to be taken with probability 1, there is no transfer scheme τ which is Pareto improving. ■

REFERENCES

1. D. Abreu and A. Sen, Subgame perfect implementation: A necessary and almost sufficient condition, *J. Econ. Theory* **50** (1990), 285–299.
2. R. Aumann, Acceptable points in general cooperative n -person games, *Ann. Math. Stud.* **40** (1959), 287–324.
3. D. Bernheim, B. Peleg, and M. Whinston, Coalition proof Nash equilibria I. Concepts, *J. Econ. Theory* **42** (1987), 1–12.
4. B. Dutta and A. Sen, Implementation under strong equilibrium: A complete characterization, *J. Math. Econ.* **20** (1991), 49–67.
5. J. Farrell and E. Maskin, Renegotiation in repeated games, *Games Econ. Behav.* **1** (1989), 327–360.
6. H. Itoh, Coalitions, incentives and risk sharing, *J. Econ. Theory* **60** (1993), 410–427.
7. S. Li, Far-sighted strong Nash equilibrium and oligopoly, *Econ. Lett.* **40** (1992), 39–44.
8. C.-t. A. Ma, Unique implementation of incentive contracts with many agents, *Rev. Econ. Stud.* **55** (1988), 555–572.
9. C.-t. A. Ma, J. Moore, and S. Turnbull, Stopping agents from cheating, *J. Econ. Theory* **46** (1988), 355–372.
10. E. Maskin, Implementation and strong Nash equilibrium, in “Aggregation and Revelation of Preferences” (J. J. Laffont, Ed.), North Holland, New York, 1979.
11. D. Mookherjee, Optimal incentive schemes with many agents, *Rev. Econ. Stud.* **51** (1984), 433–446.
12. A. Rubinstein, Strong perfect equilibrium in supergames, *Int. J. Game Theory* **9** (1979), 1–12.
13. R. Selten, Re-examination of the perfectness concept for equilibrium points in extensive games, *Int. J. Game Theory* **4** (1975), 25–55.