ELSEVIER

Journal of Public Economics 59 (1996) 117–136

JOURNAL OF PUBLIC ECONOMICS

A prisoner's dilemma model of collusion deterrence

Fred Kofman[a], Jacques Lawarrée[b,c,*]

[a]*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[b]*Department of Economics, University of Washington, Seattle, WA 98195, USA*

[c]*ECARE, University of Brussels, 1050 Brussels, Belgium*

Received October 1991; final version received August 1994

Abstract

We examine a hierarchy formed by a principal, a supervisor and an agent, wherein the supervisor and the agent can collude. We consider a case where collusion-free supervisors are not available. We demonstrate first that it is easy for the principal to deter collusion by introducing a second supervisor and designing a mechanism similar to the prisoner's dilemma so that the two supervisors control each other. Since it could prove too costly for the principal to send two supervisors, a new question arises: whether it would be possible to deter collusion by sending the second supervisor with a probability less than one. We find that under reasonable assumptions on the size of rewards and punishments, the principal can prevent collusion only by 'creating' a new type of supervisor through *sometimes* informing the second supervisor of his position.

Keywords: Collusion; Hierarchies; Monitoring; Auditing

JEL classification: D73; D82; L22

1. Introduction

Who polices the police? This question has troubled mechanism-designers ever since the early days of the Roman Empire. The problem with endowing

* Correspondence to: Professor J. Lawarrée, Department of Economics, University of Washington, Seattle, WA 98195, USA. Tel: 1-206-543-5632; Fax: 1-206-685-7477; E-mail: lawarree@u.washington.edu.

SSDI 0047-2727(94)01483-3

certain people with the power to impose penalties on others is that they might use this power for a purpose other than the intended one. If they are self-interested (as economic agents are assumed to be), they will use their privileged status for their own benefit, which may differ from that of the mechanism-designer.

Many instances of this kind of behavior have been recorded: corruption of fire, health and custom inspectors; police officers, tax auditors, and regulators being bribed or 'captured'. Despite the commonplace occurrence of bribery, economic theory has not been particularly concerned with it. The literature on contract theory has considered the use of supervisors or auditors in incentive schemes, but it has generally assumed that they are simple monitoring devices. That is, these agents are not strategic, but act in the principal's interest; they are a kind of 'third arm' for the principal.

This approach overlooks an important aspect of the design of mechanisms which do use supervisors. Whenever supervisors can manipulate evidence and affect payments to third parties, they will do so in their interests. This strategic behavior constrains the set of incentive compatible mechanisms available to the designer. In this paper we explore some strategies to deal with self-interested supervisors.

Three kinds of problems may arise from the divergence between the goals of the principal and those of the supervisor. (1) If the supervisor needs to spend (unwanted) effort to find out compensation-relevant information about the agent, he may shirk and report inaccurately. (2) If the supervisor and the agent can jointly manipulate compensation-relevant information and can write self-enforcing side-contracts, they may manipulate their information to play cooperatively against the principal. (3) If the supervisor can manipulate, by himself, compensation-relevant information about the agent, he may frame and blackmail the agent.

Baiman et al. (1987) deal with the first problem in an auditing model. It seems that the third problem (framing) has yet to be studied by economists. Our paper analyzes the problem of side-contracts.

Our concern is to find a way to prevent collusion in hierarchies of self-interested agents and supervisors. The key idea is to use a second supervisor to monitor the first one. If the principal decides to use two supervisors, the question arises as to who will monitor the second supervisor. Collusion deterrence depends on the probability of detection, so if the second supervisor is not monitored he will collude and lose all his effectiveness for the principal. This reasoning leads inevitably to an infinite regress; we need a third supervisor to monitor the second, a fourth to monitor the third, and so on.

We will show that it is possible to design a system of rewards and punishments à la prisoner's dilemma so that the two supervisors police each other. Even though double-checks are a good idea, they are costly. The cost

of sending two supervisors may cause such an arrangement to be sub-optimal. We ask, then, whether it is possible to deter collusion by sending the second supervisor with probability less than one.

We conclude that, under reasonable assumptions on the size of rewards and punishments, the principal can achieve truthful reporting only by 'creating' a new type of supervisor. When sending the two supervisors sequentially the principal cannot stop collusion if he tells them either always or never whether they are the first or the second supervisor. However, by *sometimes* informing the second supervisor of his position and not telling the agent whether the second supervisor is informed, he can effectively stop collusion. The intuition behind this result is that the second supervisor, *when informed about his position*, will require a bribe unprofitable for the uninformed agent to pay, given that a bribe has already been paid to the *first uninformed supervisor*. In other words, the second supervisor will never collude when he knows his position *and* when the first supervisor does not know his position.[1] By introducing the imperfect information the principal 'creates' this new type of supervisor and is able to deter collusion.

The mechanism we propose is an example of Bayesian Perfect implementation. That is to say, we trim the set of equilibria of the game defined by the principal using the Bayesian Perfect Nash criteria.

There is a variety of settings in the agency literature wherein the principal gains from withholding information from the agent (Maskin and Tirole, 1990). Our analysis extends this idea to a hierarchical setting where the principal garbles his communication with the supervisor as opposed to the agent. In contrast to Maskin and Tirole, our principal does not design a contract that maximizes ex post information asymmetry between him and the agent and preserves his private information. In our framework, the principal has no private information; instead, by creating a hidden randomization—which does not affect the exogenous parameters of the model—the principal creates the *source* of his private information.

Tirole (1986) was the first to study the phenomenon of bribes in a hierarchical contract involving a principal, a supervisor and an agent. However, Tirole rules out the possibility of adding a second supervisor.

In Kofman and Lawarrée (1993), we derive the optimal contract when both an internal and an external auditor are available. However, the external auditor never colludes by assumption. We study the effect of collusion on the agent's incentives to exert an effort. This paper, however, rules out the existence of a collusion-free supervisor and models instead two identical supervisors, focusing on cross-checking mechanisms.

Laffont and Martimort (1994) also investigate the simultaneous use of two collusive supervisors. They show that information per se introduces

[1] If they both know their position, collusion cannot be prevented.

increasing returns in the benefits of side-contract. By duplicating auditors, the principal can reduce their information and their discretion, and, therefore, improve expected welfare.

Tirole (1992) surveys the recent literature on collusion in organizations.

The paper is organized as follows. In Section 2 we model our problem as a game theoretical situation, discuss reasonable assumptions about the magnitude of penalties and rewards, and we present a simple, collusion-proof mechanism. In Section 3 we explore the possibility of sending the second supervisor with a probability less than one and show that a simple model yields counter-intuitive results. Section 4 presents a resolution of this problem and restores our initial intuition. Finally, Section 5 gathers our conclusions.

2. Description of the model

We consider a vertical structure represented by a three-layer hierarchy: principal–supervisor–agent. The principal owns a productive technology, but lacks the skills or the time necessary to operate it and must hire an agent for that purpose.[2] The agent is the productive unit. The principal also lacks the knowledge to supervise the agent. He can hire supervisors whose only role is to audit the agent.[3] We assume, in addition, that supervising does not require any effort from the supervisors (this avoids the moral hazard problem) but is costly to the principal: its cost is equal to the reservation wage of the supervisors (W). All players are risk neutral.

The agent's ability to perform depends on a characteristic unobservable to the principal. This characteristic (or type) is the agent's private information and determines his productivity. We assume that the agent can only be of two types: high productivity or low productivity. In the second-best contract without supervisors the high productivity agent obtains an informational rent π. (For an underlying structure yielding this result, see Baron and Myerson, 1982, or Laffont and Tirole, 1986.) We assume that the agent has

[2] We do not allow the principal to sell the firm. We thereby limit ourselves to contracts which maintain the principal as the residual claimant of the vertical structure.

[3] The use of supervisor(s) may not be the only way for the principal to achieve better control of his agent. In some cases, it is not even feasible as in the case of relationships between doctors and patients, lawyers and clients, advisors and Ph.D. students. The principal may also duplicate the agents to get more information. This method, however, may be very inefficient. A typical example is the regulation of a private firm (agent) characterized by increasing returns to scale. The existence of competition can reduce the incentive problem (see Hart, 1983; Scharfstein, 1988; and Hermalin, 1992) but some of the benefits of scale economies will be wasted. Our model studies the efficiency of using a third party (supervisor) to lessen the information asymmetry problem. It is then assumed that other institutional arrangements are not feasible.

limited liability and the contract must award him a non-negative payoff in any state of the world.

To reduce the informational rents of the high productivity agent, the principal can employ a self-interested supervisor at cost W. We assume that the supervisor cannot buy the right to audit. He can be subject, though, to negative transfers if he is caught lying. The supervisor learns the agent's private information without mistakes and obtains verifiable evidence. His report to the principal, however, can differ from his observation. If he can gain by manipulating the report and the agent agrees, he will do so. We require that the agent collaborates in manipulating the information to avoid cases in which the auditor can 'frame' the agent. The supervisor can gain by manipulating his reports because the agent can give him a conditional side-transfer (a bribe, B). The agent may share his rent π to get the auditor to present false information to the principal.

To prevent collusion, the principal could match the agent's bribe with a reward R. This solution does not improve the principal's payoff. Since the agent will lose π if reported to the principal, he will be willing to pay up to π to the auditor to present a false report. To discourage the auditor from doing that, the principal will have to match the bribe and pay π to the auditor when his report extracts the agent's rents. This is a 'bounty-hunter' scheme where the auditor obtains all the informational rents from the agent (see Kofman and Lawarrée, 1993).

Another strategy to prevent collusion is to hire a second auditor at cost W. This auditor is similar to the first. He is self-interested, learns perfectly the agent's private information, obtains verifiable evidence, and can manipulate this evidence with the help of the agent to give the principal a false report. As the first auditor, he cannot pay for the right to audit but can be subject to negative transfers if caught lying. The way in which a supervisor can be caught lying is that the two reports disagree. The truth-teller will have evidence to verify his report while the liar will not. When this is the case, the principal can apply a non-pecuniary punishment (P) on the lying supervisor. When two supervisors participate in the contract, the reward for the principal is not only from extracting the rents from the agent but also from potentially uncovering the false report of the other supervisor.

We are interested in situations where the principal uses the auditors, therefore we will assume that their cost is sufficiently low, i.e. $2W < \pi$.

We assume that the agent has all the bargaining power in his negotiation with the auditors. He is the only one who can commit to a side-transfer so he can make a conditional take-it-or-leave-it offer. He can offer any bribe he wants, but he will never offer bribes that add to more than his rent. (His Nash threat payoff is zero, so he will not go below that.)

We assume that the punishment to the colluding supervisors (P) cannot exceed $\overline{P}$. $\overline{P}$ is the modeler's reflection of social practices. We want to determine the minimum value of $\overline{P}$ which could prevent collusion.

We assume that the reward for the auditors (R) is lower than the rent π. If R exceeded π, the agent and the supervisor might collude and share the surplus of the coalition $R-\pi$. Note that by reporting truthfully the supervisor allows the principal to recover π, then it is reasonable to assume that R must be bounded above by π. We also assume that the reward is paid only to the supervisor uncovering collusion between the agent and the supervisor. Therefore, a supervisor reporting detrimental information about the agent does not necessarily collect a reward. We have analyzed a model where this assumption does not hold elsewhere (Kofman and Lawarrée, 1993).

Summarizing: the timing of the game is:

(1) Nature draws a type for the agent. The high productivity type agent obtains a rent of $\pi > 0$; the low productivity agent obtains no rent.

(2) The principal sends the two supervisors simultaneously under a contractual agreement that specifies transfers as a function of their reports. When the agent is high productivity, the transfers would be:

	High prod. report	Low prod. report
High prod. report	0, 0	$R, -P$
Low prod. report	$-P, R$	0, 0

(If the auditors' reports differ, the principal will reward the truth-telling and punish the liar.)

(3) Both supervisors observe the agent's type.

(4) The agent can commit to side-transfers to the supervisors conditional on their reports.

(5) Both supervisors report simultaneously.

(6) Transfers and side-transfers are realized.

A simple solution involving a prisoner's dilemma

When the principal sends both supervisors simultaneously, if $\overline{P} > \pi/2$, he can make the two supervisors play a prisoner's dilemma. Each supervisor can choose between reporting truthfully or lying. The payoff matrix is:

		Supervisor 2	
		Report truth	Lie
Supervisor 1	Report truth	0, 0	$R, B-P$
	Lie	$B-P, R$	B, B

To guarantee that the outcome (report truth, report truth) is a Nash equilibrium, we need $P \geq B$. This condition is easily verified since $B \leq \pi/2$ (by the agent individual rationality constraint) and since P can be greater than $\pi/2$ ($\overline{P} > \pi/2$ and $P \leq \overline{P}$).[4]

To guarantee that the outcome (lie, lie) is *not* a Nash equilibrium, we need $R > B$. Since $B \leq \pi/2$, $R > \pi/2$ will do the job while respecting the principal's budget constraint ($R \leq \pi$).

Therefore, if those two conditions are satisfied ($P \geq B$ and $R > B$), the principal always gets a truthful report. This mechanism seems to be extremely powerful. The only (mild) assumption we need to make is that the punishment imposed on a supervisor who accepts a bribe be slightly higher than the bribe.

If a prisoner's dilemma is so efficient, one might wonder why this type of mechanism is not observed more frequently in the real world. Restrictions on the values of P or R do not seem to be the cause. Rather, the cost of doubling the supervisory function appears to be a more serious problem.[5] Such an increase in the number of regulatory agencies or IRS auditors, for instance, might not be financially feasible. In that case, the interesting question is whether a collusion-free outcome will remain an equilibrium when the principal sends a second supervisor with some probability (call it γ) less than one. Intuition suggests that, if P can be increased, γ can be decreased proportionally. And, indeed, casual observation of the real world shows that a supervisor caught accepting a bribe suffers a punishment much higher than was the bribe itself. The limited financial liability of the supervisors can easily be overcome by using non-monetary punishments, ranging from loss of face to imprisonment.

In the next section we study a game where the supervisors are not sent simultaneously to audit the agent. We explore the possibility of sending the second one with a probability less than one when P is allowed to grow.

3. Sequential sending of supervisors

Let us call our two potential supervisors S_a and S_b. Note that the principal is completely indifferent between sending either supervisor in the first place. Therefore, let us say, without loss of generality, that he sends each with

[4] Indeed, it is reasonable to assume that the maximum punishment can be higher than the bribe.

[5] If this game is repeated, collusion is also more likely (see Kreps et al., 1982). Also, as in any prisoner's dilemma, communication between the two supervisors must be prevented. This assumption seems reasonable when the principal has a very large pool of supervisors available (government, large corporation, etc.).

probability 1/2. We also assume that the principal does not tell the supervisors whether they are the first or second. More generally, we will call ξ the probability of telling the second supervisor his position. So, here, we assume that $\xi = 0$.

At this point, it is useful to recall the timing of our game.

Assume that Nature has drawn a type of agent such that this agent can earn a positive rent ($\pi > 0$). The type is the agent's private information. (If Nature draws a type of agent such that $\pi = 0$, the timing is similar, but no bribing occurs.)

(1) The principal randomizes and sends the first supervisor who observes $\pi > 0$.

(2) The agent offers a bribe B_1 to the first supervisor. The agent can commit to B_1. The principal cannot observe B_1.

(3) (a) If the supervisor refuses the bribe, he reports that π is positive. The agent gets no rent. End of the game. (b) If the supervisor accepts the bribe, he receives B_1 and reports $\pi = 0$.

(4) The principal sends the second supervisor with probability γ, which is common knowledge.

(5) The agent offers a bribe B_2 to the second supervisor. The agent can commit to B_2. The principal cannot observe B_2.

(6) (a) If the second supervisor accepts the bribe, he reports that $\pi = 0$. The two supervisors keep their bribes and the agent collects π. End of the game. (b) If the second supervisor refuses the bribe, he reports $\pi > 0$ and collects R. The first supervisor keeps his bribe, but is punished with P. The agent loses the bribe to the first supervisor and does not collect π. The principal collects π and pays the reward to the second supervisor.[6] End of the game.

The equilibrium concept we will use is the Perfect Bayesian Equilibrium (Fudenberg and Tirole, 1991). Loosely speaking, the strategies chosen by each player must be their best response to the other player's strategy, and their posterior beliefs are derived from their prior beliefs using Bayes' Rule. In this game, the agent must choose the amount of the bribes ($B_1 \geq 0$ and $B_2 \geq 0$) and the supervisors must decide whether to accept the bribe. This game is played conditional upon a fixed strategy of the principal. This strategy is characterized by the parameters γ, R, P and ξ. The fixed strategy of the principal should be feasible, i.e. belong to a strategy set described by means of the following constraints: $R \leq \pi$, $0 \leq \gamma \leq 1$, $P \leq P$, $0 \leq \xi \leq 1$.

When a supervisor must decide whether to accept or reject a bribe, it is very important for him to know if he is the first or the second supervisor. Suppose, for instance, that he knows that the first supervisor has already

[6] Remember that we assume that the principal does not collect the punishment P (because it takes the form of a jail term, for instance).

accepted a bribe. In that case, he would simply compare the bribe offered by the agent with the reward he could get from the principal by denouncing the first supervisor. However, if he knows he is the first supervisor, his action will depend on his beliefs about the likelihood of the second supervisor accepting the bribe.

The probability of being called as the first supervisor is 1/2. Recall that γ is the probability of the principal sending a second supervisor when the first has reported that $\pi = 0$. Call β the probability that the first supervisor colludes.

Then, the probability of being called as the second supervisor is $(1/2)\beta\gamma$ and the probability of being called as a supervisor is $(1/2)(1 + \beta\gamma)$.

It is useful to remember that we assume that auditing is perfect. When a supervisor is sent to audit the agent, he immediately knows the agent's characteristic.

Now, using Bayes' Law, a player can compute the probability of being the first supervisor given that he has been called. For instance, A could compute:

$$\text{Prob (A is 1st}|\text{A was called)} = \frac{\text{prob (A is called 1st)}}{\text{prob (A is called as supervisor)}}.$$

The probability of being the first supervisor given 'called' is then $(1/2)/[(1 + \beta\gamma)/2] = 1/(1 + \beta\gamma)$ and the probability of being the second supervisor given 'called' is $\beta\gamma/(1 + \beta\gamma)$.

Note that these two probabilities are intuitive results. Suppose $\beta = 0$ (the other supervisor never accepts the bribe); then, when a supervisor is called he knows he cannot be the second supervisor since the first supervisor would already have refused the bribe and denounced the collusion.

In order to find an equilibrium, we now have to compute the expected payoff of a supervisor (recall that the two supervisors are identical) who is considering whether to accept or to reject a bribe:

$$\begin{aligned}\text{expected payoff (refuse)} &= \frac{W}{1 + \beta\gamma} + \frac{\beta\gamma}{1 + \beta\gamma}(R + W) \\ &= W + R\frac{\beta\gamma}{1 + \beta\gamma}. \qquad (1)\end{aligned}$$

The intuition is as follows. If he is the first player, then, by refusing the bribe, he will not get any payoff in excess of his wage W. However, if he is the second player, he can also get the reward (R) since the first supervisor has cheated.

$$\text{expected payoff (accept)} = \frac{1}{1+\beta\gamma}\{W + (1-\gamma)B_1 + \gamma[\beta B_1 + (1-\beta)(B_1 - P)]\} + \frac{\beta\gamma}{1+\beta\gamma}(W + B_2) = W + \frac{B_1 + \beta\gamma B_2}{1+\beta\gamma} - P\left(\frac{\gamma(1-\beta)}{1+\beta\gamma}\right). \quad (2)$$

If he is the first player, he will get the bribe (B_1) if the second supervisor also accepts the bribe. However, if the second supervisor refuses the bribe, the first player will be punished (P) by the principal. On the other hand, if he is the second player, he is certain to get the bribe.

The decision rule of both supervisors is to compare (1) and (2), choosing the higher one. A supervisor will refuse the bribe if

$$W + R\frac{\beta\gamma}{1+\beta\gamma} > W + \frac{B_1 + \beta\gamma B_2}{1+\beta\gamma} - P\left(\frac{\gamma(1-\beta)}{1+\beta\gamma}\right), \quad (3)$$

which implies

$$R\beta\gamma > (B_1 + \beta\gamma B_2) - P\gamma(1-\beta). \quad (4)$$

We will restrict our attention to the case where, in equilibrium, β equals zero or one. Each supervisor is sure that the other supervisor either will never collude or will always collude. Other possible equilibria would require the use of mixed strategies by both players. This restriction rules out semi-separating equilibria as well.

In pure strategies, we must also consider two possible kinds of equilibria. If $B_1 = B_2$, we have a pooling equilibrium. If $B_1 \neq B_2$, we have a separating equilibrium. For this type of equilibrium we must consider whether the agent can credibly reveal or signal to each supervisor his position. Note that, in this case, true signaling must be optimal ex post as we assume that the agent cannot commit to truthful revelation of the supervisors' position. We begin by considering the case of a pooling equilibrium.

Since this game has several equilibria, our approach is to find out if the principal has a feasible fixed strategy such that no bribe is accepted in equilibrium.

Definition. For a given fixed strategy of the principal, the game's equilibrium set is called collusive (or a collusion equilibrium set) if it contains at least a collusion equilibrium, i.e. an equilibrium in which a bribe is accepted by a supervisor with positive probability.

3.1. Pooling equilibrium

A first equilibrium can be found when prior beliefs (with respect to β) are zero for both supervisors: each supervisor thinks the other will never accept a bribe. With $\beta = 0$ an equilibrium would require:

$$B < P\gamma \,. \tag{5}$$

In other words, if the expected punishment is higher than the bribe, a situation where nobody colludes is a Nash equilibrium. Note that if P is not bounded above, we can get a collusion-free equilibrium with γ (the probability of sending a second supervisor) arbitrarily close to zero.

An equilibrium where both supervisors accept the bribe (i.e. $\beta = 1$) will occur if

$$B > R\frac{\gamma}{1+\gamma}\,. \tag{6}$$

In order to apprehend the intuition behind this formula, consider the case where $\gamma = 1$ (when the principal receives a report from his first supervisor saying that the agent has reported his true characteristic, he always sends a second one). In this case, the condition is $B > R/2$. Each supervisor compares the bribe with his expected reward, equal to $R/2$ since he has one chance in two to be the second supervisor.

The above two equilibria are not incompatible. Conditions (5) and (6) could be simultaneously satisfied. The outcome would depend only on the prior beliefs of both supervisors. However, for the supervisors, the equilibrium where $\beta = 1$ is Pareto dominant. They would be better off since $W < W + B$. This would then be the expected equilibrium outcome. The question becomes: Can the principal find a feasible[7] fixed strategy to prevent the collusion pooling equilibrium?

We obtain the following proposition:

Proposition 1. When $\xi = 0$, (i) the principal cannot prevent a collusive (pooling) equilibrium set. (ii) Moreover, the collusion pooling equilibrium Pareto dominates (for the agent and the supervisors) the other non-collusion pooling equilibrium.

Proof. (i) Suppose that the collusion equilibrium can be prevented. In a

[7] With no restriction on the values of P and R, the task of the principal would be easy. He could choose $R \to \infty$, $P \to \infty$ and $\gamma \to 0$ (P going to infinity faster than γ goes to zero, and R going to infinity more slowly than γ goes to zero). Then, the only possible equilibrium is $\beta = 0$ (nobody colludes), and, since $\gamma \to 0$, the principal almost never wastes money sending a second supervisor.

pooling equilibrium, the bribe has to be the same for both supervisors, $B_1 = B_2 = B$. The agent will then accept to pay $B(1+\gamma) \leq \pi$ or $B \leq \pi/(1+\gamma)$. So the maximum bribe that the agent is willing to pay is $B^{\max} = \pi/(1+\gamma)$. Remember, to collude, a supervisor must get at least $B^{\min} = R[\gamma/(1+\gamma)]$, and notice that $B^{\min}$ is independent of P. Even if the punishment is large, it does not affect the decision to collude.

The collusion equilibrium will be avoided if $B^{\max} < B^{\min}$, i.e. if

$$\frac{\pi}{1+\gamma} < R\frac{\gamma}{1+\gamma} \Leftrightarrow \frac{\pi}{R} < \gamma \leq 1.$$

However, since it is assumed that $R \leq \pi$, then $\pi/R \geq 1$; so we have a contradiction.[8]

(ii) The collusion equilibrium is Pareto dominant for the agent and the supervisors since the other (non-collusion) equilibrium drives all the players down to their reservation utility (0 for the agent and W for the supervisors). □

The result of Proposition 1 is somewhat surprising since it still holds when an infinite punishment is available.

3.2. Separating equilibrium

A separating equilibrium arises when the supervisors are informed about their position. This situation could arise either (i) because the agent informs the supervisors by offering different bribes to the first and second supervisors; or (ii) because the principal informs the supervisors ($\xi = 1$).

Case (i): the principal does not inform the supervisors about their position ($\xi = 0$).

In this case a separating equilibrium will exist only if the agent informs the supervisors about their position. However, in that case, no collusion separating equilibrium exists, as shown in the following proposition. Therefore, it is not in the interest of the agent to inform the supervisors.

Proposition 2. When $\xi = 0$, a collusion separating equilibrium does not exist.

Proof. In a equilibrium where $\beta = 1$, bribes must satisfy individual rationality constraints for the two supervisors: $B_1 \geq 0$ and $B_2 \geq R$. In addition, they must satisfy incentive constraints for the agent. This requires that the agent

[8] When the principal chooses $\gamma = 1$ and $\pi = R$, a supervisor with prior beliefs $\beta = 1$ would be indifferent between accepting the bribe or refusing it. Strictly speaking, collusion would still be a Nash equilibrium of this game. As we stated before, it is also much more likely because it is Pareto dominant for the supervisors.

has no incentive to tell the first supervisor that he is second: $B_1 \leq B_2$. Also it requires that the agent has no incentive to tell the second he is the first: $B_2 \leq B_1$. Together, these (separating) conditions imply $B_1 = B_2$ and $B_1 + B_2 \geq 2R$. It is then cheaper for the agent to induce a pooling equilibrium with $B_1 = B_2 = B = R\gamma/(1+\gamma)$. □

Therefore, if the principal does not inform the supervisors about their position ($\xi = 0$), the agent will not either, and a pooling equilibrium will result. Since the pooling equilibrium is likely to involve collusion, the principal may decide to inform the supervisors ($\xi = 1$) if this helps him deter collusion.

Case (ii): the principal does inform the supervisors about their position ($\xi = 1$).

It is straightforward to establish

Proposition 3. When $\xi = 1$, the equilibrium set is collusive, i.e. the principal cannot prevent a collusion separating equilibrium.

Proof. The second supervisor colludes if $B_2 \geq R$. The first supervisor colludes if $B_1 \geq 0$ and $B_2 \geq R$. The agent finds it profitable to bribe the first supervisor if $B_1 + \gamma B_2 \leq \pi$ and the second supervisor if $B_2 \leq \pi - B_1$. This second condition is stronger than the first one. To prevent collusion, it is necessary that $R > \pi$, which is not feasible for the principal. □

Propositions 1, 2 and 3 yield counterintuitive results. Under assumptions we found reasonable (mainly $R \leq \pi$), the principal could not prevent a collusive equilibrium set,[9] whatever the punishment was. This surprising result stems from the fact that, in a collusion equilibrium, the supervisors do not have to worry about the punishment since they will never have to bear it. In the next section we show that the principal can overcome this problem by introducing some asymmetric information. The principal can indeed choose the probability of telling the second supervisor whether or not he is second strictly between zero and one, i.e. $\xi \in (0, 1)$.

Note that the result of Propositions 1, 2 and 3 (impossibility of deterring collusion) is a knife-edge result. Letting the principal use an $R > \pi$ would break it and imposing an $R < \pi$ would reinforce it. Our point is, therefore, not to claim the generality or robustness of this result, but simply to highlight the dramatic effect of introducing imperfect information.

[9] Remember that a collusive equilibrium set may also include a non-collusive equilibrium.

4. The solution

Consider the scheme where with probability ξ the principal informs the second supervisor that he is such $(\xi \in [0, 1])$.[10,11] $\xi = 0$ implies the uninformed supervisors model and $\xi = 1$ the informed supervisors model.

Proposition 4. By choosing the probability of telling the second supervisor whether he is second or not strictly between zero and one, the principal can prevent the collusion equilibrium (i) if $R > \pi/(1 + \gamma - \xi\gamma)$, *which prevents the separating equilibrium, and (ii) if*

$$R > \max\left\{\frac{\pi}{1+\gamma}, \frac{\pi - (\pi + P)\gamma\xi}{\gamma(1-\xi)}\right\},$$

which prevents the pooling equilibrium.

The rest of this section will prove Proposition 4. Subsection 4.1 deals with the separating equilibrium and Subsection 4.2 with the pooling equilibrium.

4.1. The separating equilibrium

To deter the collusive separating equilibrium, we need to have that $R > \pi/(1 + \gamma - \xi\gamma)$.

Proof. We look for necessary conditions for a separating equilibrium to prevail. Clearly, B_1 and B_2 must be individually rational for the supervisors, i.e.

$$B_1 \geq 0 \quad \text{and} \quad B_2 \geq R\,. \tag{IR}$$

Next, B_1 and B_2 must satisfy incentive compatibility constraints for the agent. More precisely, there are two cases here: either $B_1 \geq R$ or $B_1 < R$.

(i) $B_1 < R$. The agent has no incentive to tell the first supervisor that he is the second if

$$\pi - B_1 - \gamma B_2 \geq \pi - B_2 - \gamma B_2\,, \tag{IC1}$$

which is obviously equivalent to $B_2 \geq B_1$.

[10] The first supervisor is never told.

[11] At this point, we assume that the principal can commit to such a scheme. However, since we will show that the first supervisor never accepts a bribe in equilibrium, the principal is indifferent between telling the second supervisor about his position or not.

The agent has no incentive to tell the second supervisor that he is the first (taking into account the probability that the second supervisor can turn the agent in) if

$$\pi - B_1 - B_2 \geq \pi - B_1 - (1-\xi)B_1 - \xi\pi\,, \tag{IC2}$$

which is equivalent to

$$B_2 \leq (1-\xi)B_1 + \xi\pi.$$

Since $B_2 \geq R$, then necessarily

$$\frac{R-\xi\pi}{1-\xi} \leq B_1,$$

and finally

$$\gamma R + \frac{R-\xi\pi}{1-\xi} \leq B_1 + \gamma B_2.$$

Thus, if

$$\pi < \gamma R + \frac{R-\xi\pi}{1-\xi}\,, \tag{*}$$

the agent's expected payoff is negative, meaning that the strategy (B_1, B_2) is not optimal for the agent, who should not induce such a separating equilibrium. Condition (*) is equivalent to $R > \pi/(1+\gamma-\xi\gamma)$.

(ii) $B_1 \geq R$. Conditions (IC1) and (IC2) then boil down to $B_1 \geq B_2$ and $B_2 \geq B_1$, since the second supervisor accepts the bribe, even if he is informed by the principal. Thus, $B_1 = B_2$ and $B_1 + B_2 \geq 2R$. Then, if

$$R > \pi/2\,, \tag{**}$$

such a strategy is not a best response of the agent.

Condition (*) is stronger than (**). □

4.2. *The pooling equilibrium*

Two types of strategy profiles which could be candidates for a collusion pooling equilibrium have to be considered:

(1) The bribe to both supervisors is R:$B_1 = B_2 = R$.

Bribing will be unprofitable for the agent if $R(1+\gamma) > \pi$ or $R > \pi/(1+\gamma)$.

(2) The bribe to both supervisors is B:$B_1 = B_2 = B < R$. Note that B must be greater than or equal to $B^{\min}$, the minimum bribe required by an uninformed supervisor, otherwise no supervisor will accept the bribe. The ap-

pendix computes $B^{\min}$ and shows that setting $R \geq [\pi - (\pi + P)\gamma\xi]/[\gamma(1 - \xi)]$ makes it unprofitable for the agent to offer any $B \geq B^{\min}$.

In summary, the conditions to prevent collusion are:

[1] $R > \pi/(1 + \gamma - \xi\gamma)$ (to prevent the separating equilibrium).

[2] $R > \pi/(1 + \gamma)$ and $R > [\pi - (\pi + P)\gamma\xi]/\gamma(1 - \xi)$ (to prevent the pooling equilibrium).

Notice that conditions [1] and [2] can be simultaneously satisfied with $0 < \xi < 1$, $0 < \gamma < 1$ and $R \leq \pi$ if P is large enough. On the other hand, these conditions cannot be satisfied if P is zero or ξ is zero or one.

The main results of this section can be summarized as follows. Introducing a $\xi \in (0, 1)$ has remarkable consequences in our model. While the collusion equilibrium cannot be prevented and is likely to occur when $\xi = 0$ or 1 (even with unbounded penalties), we found a more reassuring result when $\xi \in (0, 1)$. In that case, the collusion equilibrium can be prevented for sufficiently high P.

5. Conclusions

In this paper we show that the prisoner's dilemma can be a powerful tool when used to deter collusion between two supervisors with the same information. We also ask if the principal could achieve the same result when the second supervisor is sent with a probability less than one. Then we show that, under reasonable assumptions about the size of punishments and rewards, the principal could deter collusion only by 'creating' a new type of supervisor. By *sometimes* revealing to the second supervisor his position, the principal makes bribing unprofitable for the agent. The intuition is that a supervisor who knows he is the second also knows that the first supervisor has colluded. He can therefore receive a reward by telling the truth, and will require a bribe that the agent cannot afford to pay.

An interesting application of this model concerns organizations for which outside audits are not feasible. An example would be the Internal Revenue Service (IRS). A recent Congressional committee investigated misconduct (bribing, cover-ups, etc.) by senior managers in the IRS (US Congress, 1990). The Congresspersons recognized the difficulty of having outside auditors help in solving this problem. Indeed, IRS agents are not allowed to release tax return information outside the Treasury Department (Section 6103 of Internal Revenue code). This makes it almost impossible for an outsider to the Treasury Department to investigate any suspicion of misconduct. Our model suggests a way to deter collusion even when outside audits are not feasible.

Acknowledgements

We thank Yoram Barzel, Patrick Bolton, Richard Gilbert, Michael Katz and Richard Zeckhauser for very helpful discussions and comments. Two anonymous referees have significantly contributed to the content and presentation of the paper.

Appendix

This appendix computes the value of $B^{\min}$, the minimum bribe required by an uninformed agent to collude, and derives the conditions that make it unprofitable for the agent to bribe both supervisors with $B^{\min}$.

We consider here a case where $B^{\min} < R$.

The following lemma is useful:

Lemma. In a pooling equilibrium, if $B^{\min} < R$, the informed second supervisor will not collude.

Proof. This is obvious since the informed supervisor requires R to collude. □

Let us now find first the consistent beliefs of the supervisors when $0 < \xi < 1$. The event of being told his position is denoted by T, and not told by NT.

When a supervisor is called, he has the beliefs previously calculated about his type:

$$\text{Prob}(1\text{st}|\text{called}) = 1/(1+\beta\gamma) = \text{P}(1),$$

$$\text{Prob}(2\text{nd}|\text{called}) = \beta\gamma/(1+\beta\gamma) = \text{P}(2).$$

We now call these probabilities the prior beliefs of the supervisors.

Once he is sent to the agent, he will either be told by the principal or he will not. In any case, he will update his prior beliefs. If he is told, he knows he is the second with probability one. If he is not told, using Bayes' Rule, we can calculate:

$$\text{P}(1|\text{NT}) = \text{P}(1 \text{ and NT})/\text{P}(\text{NT}) = \text{P}(1)/\text{P}(\text{N})$$

$$= \text{P}(1)/[\text{P}(1) + \text{P}(2)(1-\xi)] = \frac{\dfrac{1}{1+\beta\gamma}}{\dfrac{1+\beta\gamma(1-\xi)}{1+\beta\gamma}}$$

$$= 1/[1+\beta\gamma(1-\xi)],$$

$$\mathrm{P}(2|\mathrm{NT}) = 1 - \mathrm{P}(1|\mathrm{NT}) = \frac{\beta\gamma(1-\xi)}{1+\beta\gamma(1-\xi)}.$$

In order to find the decision rule of the uninformed supervisor, we will compute his expected payoff when he accepts the bribe and when he refuses it.

$$\text{expected payoff (refuse)} = W + \mathrm{P}(2|\mathrm{NT})R$$
$$= W + \frac{\beta\gamma(1-\xi)}{1+\beta\gamma(1-\xi)}R,$$

$$\text{expected payoff (accept)} = W + \mathrm{P}(1|\mathrm{NT})\{(1-\gamma)B + \gamma(1-\xi)[\beta B + (1-\beta)(B-P)] + \gamma\xi(B-P)\} + \mathrm{P}(2|\mathrm{NT})B.$$

When the supervisor is the first one, two cases are possible: either the principal does not send a second supervisor (with probability $1-\gamma$) and the first supervisor gets the bribe B; or a second supervisor is sent (with probability γ). This second supervisor may be informed (with probability ξ) or uninformed (with probability $1-\xi$). If he is uninformed, he can collude with a probability β. On the other hand, our lemma has proved that, if the supervisor is informed, he will never collude in equilibrium.[12]

Plugging in the value of P(1|NT) and P(2|NT) yields:

$$\text{expected payoff (accept)} = W + B - P\frac{\gamma - \gamma\beta(1-\xi)}{1+\gamma\beta(1-\xi)}.$$

An equilibrium where the uninformed supervisors do not collude ($\beta = 0$) requires

$$\text{expected payout (refuse)}|_{\beta=0} \geq \text{expected payoff (accept)}|_{\beta=0}$$
$$\gamma \mathrm{P} \geq \mathrm{B}.$$

This condition is identical to our previous one.

An equilibrium where uninformed supervisors collude ($\beta = 1$) requires

$$\text{expected payoff (refuse)}|_{\beta=1} \leq \text{expected payoff (accept)}|_{\beta=1}$$
$$\gamma(1-\xi)\mathrm{R} \leq \mathrm{B}[1+\gamma(1-\xi)] - \mathrm{P}\gamma\xi.$$

[12] Although it is feasible for the agent to bribe the informed supervisor, it is not profitable for him to do so since the supervisor's type is unknown to the agent.

Again, setting $\xi = 0$ would yield our previous condition. The important new feature in this formula is the presence of P. Intuitively, with some positive probability ($\gamma\xi$), a second supervisor will be informed, refuse the bribe and subject the first supervisor to the punishment. Therefore, to rule out the equilibrium where the uninformed supervisor colludes, we need

$$P > B\left[\frac{1+\gamma(1-\xi)}{\gamma\xi}\right] - \left(\frac{1-\xi}{\xi}\right)R\,. \qquad [*]$$

If P is assumed to be unbounded, this condition can be fulfilled even if R is very small. The assumption of infinite punishment is, however, unsatisfactory with respect to the real world. It is therefore interesting to check when this assumption is binding in our model.

Since $B^{\min} < R$, when a second supervisor is informed, he will always report the truth. Now, both the agent and the first supervisor who accepted the bribe face a risk: the first supervisor might have to pay a punishment, and the agent might be required to refund π after he has already paid B. The new individual rationality constraint for the agent is

$$[1+\gamma(1-\xi)]B \leqslant \pi(1-\gamma\xi)\,. \qquad [**]$$

It is now possible to determine the minimum value of P which can prevent collusion. Using condition [**], we can compute the maximum bribe the agent can offer in equilibrium:

$$B^{\max} = \frac{\pi(1-\gamma\xi)}{1+\gamma(1-\xi)} < \pi.$$

Rearranging condition [*] yields the minimum bribe an uninformed supervisor would require:

$$B^{\min} = \frac{\gamma(1-\xi)R + P\gamma\xi}{1+\gamma(1-\xi)}.$$

An equilibrium where the uninformed supervisors collude with the agent will be prevented if

$$B^{\max} = \frac{\pi(1-\gamma\xi)}{1+\gamma(1-\xi)} < \frac{\gamma(1-\xi)R + P\gamma\xi}{1+\gamma(1-\xi)} = B^{\min}$$

$$\Leftrightarrow R > \frac{\pi - (\pi+P)\gamma\xi}{\gamma(1-\xi)}.$$

References

Baiman, S., J. Evans and J. Noel, 1987, Optimal contract with a utility maximizing auditor, Journal of Accounting Research 25, 217–244.

Baron, D. and R. Myerson, 1982, Regulating a monopolist with unknown costs, Econometrica 50, 911–930.

Fudenberg, D. and J. Tirole, 1991, Perfect Bayesian equilibrium and sequential equilibrium, Journal of Economic Theory 53, 236–260.

Hart, O., 1983, The market mechanism as an incentive scheme, Bell Journal of Economics 14, 366–382.

Hermalin, B., 1992, The effects of competitive pressures on executive behavior, Rand Journal of Economics 23, 350–365.

Kofman, F. and J. Lawarrée, 1993, Collusion in hierarchical agency, Econometrica 61, 629–656.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson, 1982, Repeated prisoner's dilemma, Journal of Economic Theory 27, 245–252.

Laffont, J.-J. and D. Martimort, 1994, Duplication of regulators against collusive behavior, mimeo, IDEI, Toulouse.

Laffont, J.-J. and J. Tirole, 1986, Using cost observations to regulate firms, Journal of Political Economy 94, 614–641.

Maskin, E. and J. Tirole, 1990, The principal–agent relationship with an informed principal: The case of private values, Econometrica 58, 379–409.

Scharfstein, D., 1988, Product market competition and managerial slack, Rand Journal of Economics 19, 147–155.

Tirole, J., 1986, Hierarchies and bureaucracies: On the role of collusion in organizations, Journal of Law, Economics and Organizations 2, 181–214.

Tirole, J., 1992, Collusion and the theory of organizations, in: J.-J.Laffont, ed., Advances in Economic Theory, Sixth World Congress (Cambridge University Press, Cambridge).

U.S. Congress, 1990, Misconduct by senior managers in the Internal Revenue Service, House Report 101-800, October.