Quarterly Journal of Political Science, 2016, 11: 279–312

Collusion, Blackmail and Whistle-Blowing

Leonardo Felli[1] and Rafael Hortala-Vallve[2*]

[1] *London School of Economics; lfelli@econ.lse.ac.uk*
[2] *London School of Economics; r.hortala-vallve@lse.ac.uk*

ABSTRACT

Whistle-blowing is usually regarded as a way to identify abuse and wrongdoing on the part of governments and corporations. In this paper we show how, at a micro level, whistle-blowing can be used as a designer tool to prevent opportunistic behavior, that takes the form of collusion or blackmail, on the part of members of a simple hierarchical structure.

We focus on a three-layered principal–supervisor–agent structure and show how the principal can use whistle-blowing as a way to prevent the supervisor and the agent from colluding to the detriment of the principal.

To understand our mechanism we need to explicitly define the penalty a party has to incur for walking away from a collusive agreement. Rewarding whistle-blowing creates incentives for the uninformed colluding party to walk out of the side deal and report to the principal that collusion took place. This threat clearly reduces the informed party's incentive to participate in side deals. It also serves as a potential blackmail threat between the colluding parties. However, careful use of whistle-blowing allows the principal to eliminate opportunities for blackmail.

Keywords: Collusion; blackmail; whistle-blowing; organizations; mechanism design; communication; opportunistic behavior

Whistle-blowing is an important phenomenon both in public life and in the corporate world. High profile whistle-blowers such as Edward Snowden for

*This paper partially replaces a previous version entitled "Avoiding Collusion through Discretion" (Felli and Hortala-Vallve, 2011). We are grateful to the Editor of the journal and three anonymous referees for their insightful comments. We greatly benefited from the

MS submitted on 30 April 2015; final version received 22 June 2016
ISSN 1554-0626; DOI 10.1561/100.00015060

the National Security Agency, Bradley Manning for the US Army and M.N. Vijayakumar for the Indian Administrative Service have left a lasting and sometimes controversial impression on public opinion about the power of whistle-blowing and its disciplinary role within public and corporate life. In all these cases whistle-blowing is understood as the disclosure (in the public interest) of an illegal or damaging practice to the press or other media platforms in the name of the public interest. This usually leads to the indictment of the organization that the whistle-blower is working for.[1] In this paper we explore a very different role for whistle-blowing. We explore how rewarding whistle-blowing can be used as a mechanism design tool to prevent detrimental or unwanted opportunistic behavior such as collusion or blackmail within an organization. Indeed, in the British Standards' *Whistleblowing Arrangements Code of Practice* (2008) it is stated that "... an organisation where the value of open whistleblowing is recognised will be able to deter wrongdoing". Our findings should promote the safeguarding of whistle-blowing practises: no wrongdoing should ever occur when public organizations rationally anticipate their own malpractices will reach the public eye.

In the classical paper on Congress' oversight responsibilities, McCubbins and Schwartz (1984) argue that there is one form of oversight that is usually overlooked precisely because of its passive role. The authors assert that Congress can *police-patrol* executive agencies or can also *fire-alarm oversight* by establishing a set of criteria the agency should comply with while allowing interested parties to blow the whistle when they observe any wrongdoing. In this paper we analyze in detail the interaction between both oversight capabilities and show how Congress and ultimately the Voter (either of them can be interpreted as the principal in our model) can design contracts so as to avoid any wrongdoing. More specifically, we are interested in situations in which agencies (the supervisor in our model) might be colluding with a regulated company (the agent in our model) to the detriment of the public interest.[2] Agencies gather information on the firms they regulate yet these regulated firms would sometimes prefer this information being concealed so as to avoid tighter regulation. Colluding behavior between agencies and regulated firms (to the detriment of the public interest) can be avoided at no cost by

comments and suggestions of Philippe Aghion, Richard Arnott, Giuseppe Bertola, Sandro Brusco, Francesca Cornelli, Donald Cox, Mathias Dewatripont, Peter Diamond, Antoine Faure-Grimaud, Franklin M. Fisher, Oliver Hart, Andrea Ichino, Jean-Jacques Laffont, David Martimort, John Moore, Patrick Rey, Lars Stole, Jean Tirole, Miguel Villas-Boas, Oved Yosha and of seminar participants at numerous institutions during the exceedingly long gestation of this research project.

[1] The literature on whistle-blowing is vast, see Miceli *et al.* (2008) for a complete overview of the recent developments in multiple disciplines on the issue.

[2] When the Libor fixing scandal became public in 2012, there were indications that the regulators (the Bank of England and the Financial Services Authority) might not have been free of guilt in Barclays' behavior. Similarly, the Food and Drug Administration was supposedly working together with AdvaMed (an association of medical companies) on the medical device provisions in the 21st Century Cures Act.

appropriately rewarding whistle-blowing (a type of *fire-alarm* oversight). This mechanism is costless to the public because anticipating the possibility of whistle-blowing prevents collusion or blackmailing from occurring in the first place.

The possibility of collusion between supervisors and agents is a well-known phenomenon. For example, De La O and Martel Garcia (2015) analyze the Mexican Federal government's (our principal) oversight of municipalities with local or state auditors. These auditors (our supervisor) have a mandate to audit the use of federal resources in the hands of municipal authorities (our agent). Local auditors are seen as more effective due to their access to local information yet they are also seen more prone to political pressures or collusive agreements.

The key idea of our model is very simple. Consider a stylized three-layered hierarchical structure: a principal, a supervisor and an agent. Suppose that the agent has private information on his productive ability and the supervisor only observes an imperfect signal of such ability. The principal can elicit the information from the supervisor but by doing so he introduces the possibility of collusion between the supervisor and the agent: an agent whose ability is higher than a certain value can extract larger rents when the principal believes he is a low type. However, to reach a collusive agreement the agent and the supervisor need to communicate and define the terms of the agreement. In so doing, the supervisor might become more informed about the productivity of the agent and the principal can reward the supervisor for leaking such information. If this is the case, the agent will refuse to participate in the collusive bargaining process to avoid losing the informational rent promised by the principal and collusion will be prevented. In the mechanism we present, the principal transfers the informational rent from the agent to the supervisor if the latter reports the information revealed during collusion or, equivalently, blows the whistle and reports that collusion took place. In so doing, the principal costlessly prevents harmful collusion.

The supervisor is able to exploit this additional information revealed during collusion only if she can breach the collusive agreement even if at a cost. In what follows we advance the existing literature by explicitly modeling the enforceability of collusion, in particular the cost an individual incurs by breaching a collusive agreement. The principal is then able to prevent collusion by introducing in the contract to the supervisor, a clause that compensates the supervisor for the cost incurred when breaching the collusive agreement. Notice that since the final outcome is such that the agent refuses to participate in collusion, this clause never applies in equilibrium and hence collusion is prevented at zero cost to the principal.

Allowing the supervisor to report that collusion took place, however, comes at a cost. It creates the opportunity for the supervisor to blackmail the agent by threatening to blow the whistle even in the absence of any collusion unless the agent pays part of his informational rent to the supervisor. We show that

whistle-blowing once again can solve this problem. Allowing the agent to report the supervisor's threat to the principal is enough to prevent blackmail in equilibrium. Moreover, if rewards for whistle-blowing are carefully chosen the agent's option to blow the whistle does not introduce further blackmail opportunities on the part of the agent.

1 Related Literature

Our analysis is closely related to the literature on collusion pioneered by Tirole (1986). This literature has depicted the problem as a costly one. It has modeled collusion as a fully enforceable side contract ignoring the possibility for the principal to induce one of the colluding parties to breach the side contract and report to the principal that collusion took place. This implies that the opportunity of collusion comes at a cost to the remaining parties to the contract (Faure-Grimaud *et al.*, 2003; Laffont and Martimort, 1997; Tirole, 1986).

While the early literature on collusion has analyzed collusion under hard or verifiable information,[3] the more recent literature has considered collusion-proof mechanisms in the presence of soft or unverifiable information (Laffont and Martimort 1997, 1999, 2000; Faure-Grimaud *et al.* 1999, 2000, 2001, 2003). The last four papers are the closest to ours. They consider an incentive contract involving a principal, a supervisor, and an agent and allow parties to set up fully enforceable collusive side contracts. They show that the *collusion-proof principle* (Laffont and Martimort, 1997) holds in this environment. The optimal mechanism is equivalent to a mechanism that in equilibrium does not allow the parties of the contract to engage in collusion. In other words, the optimal contract is the solution to the principal's payoff maximization problem, provided that supervisor and agent are not involved in collusion (as well as the standard individual rationality and incentive compatibility constraints). In addition, Faure-Grimaud *et al.* (2003) show that the *equivalence principle* holds, and that delegation, when interpreted as an increase in the discretionary power of members of the organization, is a way to implement the optimal collusion-proof mechanism.

The approach we present here fits into their framework, and both the collusion-proof principle and the equivalence principle apply. In other words, the optimal mechanism we construct is such that the agent and the supervisor do not engage in collusion and an increase in the parties' discretionary power is a way to implement such a mechanism. However, in contrast to these papers, in our case the principal can always avoid collusion at no cost. We differ from these papers in that we explicitly model the possibility of breaching

[3] See for example Kofman and Lawarrée(1993; 1996).

side contracts. We then allow the mechanism designer or principal to offer a mechanism that compensates the uninformed party for breaching the side deal and reporting the existence of the side deal to the principal.[4] In other words, we enlarge the message space of both the supervisor and the agent in the general mechanism that the principal offers them. When collusion takes place under asymmetric information, these enlarged message spaces serve the role of preventing any collusion and blackmail on the equilibrium path.

A number of papers have explicitly considered the effect of delegation on parties' incentives to engage in collusive agreements. Baliga and Sjöström (1998) have explored the effect of delegation in a moral hazard setting in the presence of colluding parties' limited liability. They identify the optimal delegation mechanism that achieves the outcome that is optimal in the absence of collusion.[5] This parallels our findings yet we do not rely on limited liability but rather on the colluding parties' option to breach the side contract at a cost.

Che and Kim (2006) and Celik (2009) both analyze delegation in the presence of collusion in a hidden information framework. Both papers ask whether delegation can achieve the same outcome that is optimal in the absence of collusion. While Che and Kim (2006) reach a positive answer in a very general framework, both in terms of the technology and the number of parties involved in collusion, possibly excluding some of the parties from the side deal, they do impose restrictions on the correlation between the colluding parties' information structure. Celik (2009), on the other hand, focuses on an organizational and informational structure similar to the one we consider here. He shows that delegation is not necessarily an optimal mechanism. In contrast to these papers, the mechanism we suggest provides the colluding parties with the incentive to breach the side contract and exploit the information they learn during collusion to their advantage and to the disadvantage of the other parties. This is the reason why in our framework increasing discretion is optimal.[6]

---

[4]In this respect our approach is similar to the augmented revelation mechanisms (Ma *et al.*, 1988; Mookherjee and Reichelstein, 1990) that allow a mechanism designer (the principal) to prevent strategic coordination — as opposed to collusion — among agents. See also Demsky and Sappington (1989) for a hierarchical model where coordination between the supervisor and the agent is a concern that needs to be addressed by the optimal mechanism selected by the principal.

[5]See also Kessler (2000) for a related point.

[6]Quesada (2005) explicitly models the *informed principal* problem that may arise when collusion takes place under asymmetric information. This occurs when the party offering the side contract has private information not available to the other party. In our context collusion does not lead to an informed principal situation for two competing reasons. The supervisors and the agent's information structures are nested: the supervisor knows less than the agent. We follow Laffont and Martimort (1997) and model collusion in a way that is agnostic on the extensive form of the collusion game. In other words, our results do not rely on the identity of the *(possibly informed) principal* in the side contract or on how the collusive negotiation is structured.

The literature on whistle-blowing has mainly focused on its effects on antitrust policy and crime prevention.[7] This literature has identified the optimal leniency program that may destabilize cartels or criminal organizations by identifying the optimal amount of leniency that destroys the trust of the repeated (cartel) relationship (Motta and Polo, 2003; Spagnolo, 2004) or the optimal rewards to employees for blowing the whistle to authorities on the cartel's existence (Aubert *et al.*, 2006). More recently, Ting (2008) shows that the informational advantages of whistle-blowing might be outweighed by the costs of employees exerting less effort. In a related paper, Beim *et al.* (2014) show that too much whistle-blowing decreases the informativeness of such disclosure and might yield more wrongdoing in the first place. Relatedly, Austen-Smith and Feddersen (2009) analyze the managerial choice of whistle-blowing practises for providing incentives to report violations and committing to fix such violations internally when they are privately reported. We build on this literature by focusing on whistle-blowing as a mechanism design tool of the principal and explicitly addressing the effects that rewarding whistle-blowing has on increasing the supervisor's opportunity to blackmail the agent.

Finally, there is a recent literature that focuses on the interplay of collusion, blackmail and whistle-blowing. Khalil *et al.* (2010) explore the close relationship between bribery and extortion. They show that in the presence of soft information, it is optimal to allow bribery and extortion to occur in equilibrium even if it is feasible to deter both. The reason for this is that in their paper the coalition incentive constraints are interlinked. The key difference with our analysis is the use of whistle-blowing as the tool that allows the principal to prevent both collusion and blackmail (bribery and extortion in their terminology) at no cost.[8] Leppamaki (1997, Ch. 3) also considers explicitly the interplay of whistle-blowing and blackmail in a contractual setting. While Leppamaki (1997, Ch. 3) analyzes blackmail in an incomplete-contract dynamic framework, in what follows we solve for the principal's static mechanism design problem.

2 The Model

2.1 The Parties

We model a simple three-level hierarchy. The top of the hierarchy is the residual claimant of profits: the principal (P). The bottom layer is the only level that actually produces any output: the agent (A). The intermediate level consists of a supervisor (S), who is capable of collecting information about the

[7] See Spagnolo (2008) for an extended survey of the effects of whistle-blowing on antitrust policy, and Gambetta and Reuter (1995) for its effects on prosecuting organized crime.

[8] See also Hindriks *et al.* (1999) and Polinsky and Shavell (2001) for an analysis of both corruption and extortion in a taxation and law enforcement setting, respectively.

agent's relevant characteristics. Following our example in the Introduction, one could think of the principal as being Congress; the supervisor as being a regulatory agency that gathers information on the industry's activities; and the agent as being a regulated firm whose activities are heavily influenced by the legislation Congress puts in place.

The agent is the productive unit of the hierarchy. He is endowed with a productivity parameter θ^A, $\theta^A \in \Theta^A \equiv \{\theta_1^A, \theta_2^A\}$, $\theta_2^A > \theta_1^A$. He may or may not exert a productive effort $e^A \in \mathbb{R}$, and both effort and productivity will generate an output x according to the following simple technology:

$$x = \theta^A + e^A \tag{1}$$

The agent is assumed to be risk neutral in income. His utility function is linear in income and strictly concave in effort. Disutility of effort is expressed, in monetary terms, by $d(e^A)$, where $d'(\cdot) > 0$, $d''(\cdot) > 0$, $d'''(\cdot) > 0$, for all $e^A > 0$; $d(0) = d'(0) = 0$, $d(e^A) = 0$, for all $e^A < 0$.[9] The agent's objective function is then: $w - d(e^A)$; his reservation wage is $\bar{w}$.

The supervisor has a monitoring role. She does not contribute directly to the productive process, but just provides information. If requested, she can supply the information to the principal. This is modeled by assuming that the supervisor observes a noisy signal, $\theta^S \in \Theta^S \equiv \{\theta_1^S, \theta_2^S\}, \theta_2^S > \theta_1^S$, of the agent's productivity parameter θ^A. Arguably, the agency that regulates a particular industry will never have perfect information on this industry's characteristics. This signal is *soft* or *unverifiable* information, in the sense that an outside party — the principal in particular — has no way to verify the real value of the signal besides asking the supervisor for a report and inducing, through incentives, truthful revelation. This signal is observed by the supervisor at no cost.[10]

The supervisor is risk averse. Her utility function $V(s)$ is strictly concave in the salary s: $V'(\cdot) > 0, V''(\cdot) < 0$. The supervisor has an outside option with a reservation salary $\bar{s}$.

The principal is risk neutral, and is the residual claimant of the agent's actions.

---

[9]Notice that while this is a model with both moral hazard and adverse selection the technological assumption, Equation (1), implies that for all intents and purposes this is a pure adverse selection framework (Laffont and Martimort, 2002, Ch. 7). The role of negative effort is to keep things simple and allow the high productivity agent to mimic the low productivity agent. The assumption on the third derivative of the disutility function assures concavity of the optimization problems considered later.

[10]In principle, the supervisor might have to spend a costly effort to get a strictly informative signal, as in Demsky and Sappington (1989). However, such generalization does not add much to the analysis of collusion, while considerably complicating the notation and the presentation of the model.

2.2 The Information Structure

The principal is the least informed party. His information set includes only the final levels of output x. The supervisor costlessly observes the noisy signal θ^S of the agent's productivity θ^A, and observes x. Finally, the agent has the best information structure: he knows θ^A, and can observe both the signal θ^S and x.[11] We take x to be the only *verifiable* information of the model, while θ^A is *observable* only to the agent and θ^S is *observable* to both the agent and the supervisor.

The agent's productivity θ^A and the supervisor's signal θ^S are positively, but imperfectly, correlated. Let:

$$q_i = Pr\{\theta^S = \theta_1^S \mid \theta^A = \theta_i^A\} \qquad i \in \{1, 2\} \tag{2}$$

That is, q_1 is the probability that the signal θ_1^S is correct and q_2 is the probability that the same signal is not correct. We take θ^S to be a strictly but not fully informative signal of θ^A:

$$0 < q_2 < \frac{1}{2} < q_1 < 1 \tag{3}$$

2.3 The Timing and Solution Concept

Before contracting, θ^A and θ^S are determined by nature and are the agent's and the supervisor's private information. As mentioned above the supervisor only observes the realization of θ^S while the agent observes the realizations of both θ^S and θ^A. The principal's beliefs about θ^A and θ^S are then characterized by the prior $\pi = Pr\{\theta^A = \theta_1^A\}$ and the conditional distribution q_i as in Equation (2), while the supervisor's beliefs, after observing the realization of θ^S, are:

$$\begin{aligned} p_1 = Pr\{\theta^A = \theta_1^A \mid \theta^S = \theta_1^S\} &= \frac{q_1\pi}{q_1\pi + q_2(1-\pi)} \\ p_2 = Pr\{\theta^A = \theta_1^A \mid \theta^S = \theta_2^S\} &= \frac{(1-q_1)\pi}{(1-q_1)\pi + (1-q_2)(1-\pi)} \end{aligned} \tag{4}$$

Negotiations take place in which the principal is assumed to have all the bargaining power. He proposes a *take-it-or-leave-it* contractual offer $C = (C^A, C^S)$ to both the agent and the supervisor, which specifies a schedule of compensations for them contingent on output and on the supervisor's report.

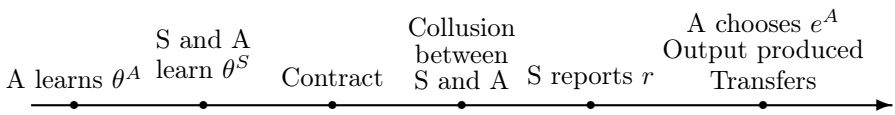
Supervisor and agent simultaneously and independently decide whether to accept or reject the principal's offer. If the agent rejects the offer, negotiation with both parties ends and the game ends. If the supervisor rejects the offer,

[11] The fact that A observes θ^S is clearly a simplifying assumption. It provides us a simple framework in which collusion may take place between the agent and the supervisor.

negotiation proceeds involving only the agent. The game then becomes a standard two-tier principal–agent problem.[12] If the principal wishes, he can make degenerate offers to the supervisor, which amount to a decision on his part to negotiate only with the agent. If both supervisor and agent accept the offer a contract is signed.

After the contract is signed, the collusive negotiation between the supervisor and the agent takes place. We provide below a general characterization of this negotiation.

At a predetermined time — between the initial contracting date and the date at which the agent produces output x — the supervisor produces a report r of her observed signal that becomes public information.[13] The agent then exerts his productive effort, the outcome of production becomes publicly observable and remunerations are paid according to the contract C.[14] We assume that all of this structure — summarized in the figure below — is common knowledge to all the parties.



2.4 A World Without Collusion or Blackmail

The revelation principle implies that, without loss of generality, we can restrict attention to a revelation game where both agent and supervisor reveal their information to the principal, who can then use this information subject to the usual incentive constraints. These incentives constraints require that the agent and supervisor reveal their information truthfully, given the contracts offered by the principal (that is, the remunerations and the effort level of the agent).[15] In the revelation game the supervisor's strategy space is the set of all possible mappings from the signal space Θ^S into her message space Θ^S.

[12] Alternatively, the principal could make a unique offer to the agent that specifies a contractual arrangement, if the supervisor accepts the principal's offer and a different arrangement if the supervisor rejects the principal's offer; nothing would change.

[13] In principle, it might be of use for the principal to ask the agent, as well as the supervisor, to report the signal θ^S. This cannot improve the Principal's utility, given that under our optimal mechanism collusion will be avoided at zero cost.

[14] We take the timing of the supervisor's report as exogenously given. This is a simplifying assumption. However, our main result — the fact that collusion can be prevented at no additional costs — suggests a reason why the principal might want to specify the timing we analyze. See Felli (1990, Ch. 2, Sec. 6) for a discussion of the case in which the timing of the supervisor's report is endogenous.

[15] See Chapter 7 of Laffont and Martimort (2002) for a detailed discussion on how in our case the adverse selection and moral hazard problem simplify to a pure adverse selection problem using the revelation principle.

Similarly, the agent's strategy space is the set of all possible mappings from the space of the productivity parameters Θ^A into the message space Θ^A. The mechanism the principal offers specifies a salary for the supervisor, a wage for the agent, and an output as functions of the agent's and supervisor's reports:

$$C = [s(\hat{\theta}^A, \hat{\theta}^S), w(\hat{\theta}^A, \hat{\theta}^S), x(\hat{\theta}^A, \hat{\theta}^S)]$$

If only the agent accepts the principal's offer then the mechanism boils down to a standard principal–agent problem.[16]

In what follows we focus on the Perfect Bayesian Equilibria (PBE) of this game (Fudenberg and Tirole, 1991): a contract (remunerations and effort level in each state of the world), a revelation mapping for both the agent and the supervisor (a mapping from their private information to the message revealed to the principal) and a belief system by the principal such that all strategies are sequentially rational given the belief system and the belief system is consistent, wherever possible, given the strategy profile.

When there is no scope for collusion, both agent and supervisor are *honest* in the sense that neither of them engage in collusion or blackmail, our simple structure allows the risk neutral principal to pay a constant salary to the risk averse supervisor

$$s(\hat{\theta}_i^S, \hat{\theta}_j^A) = \bar{s} \quad \forall i, j \in \{1, 2\} \tag{5}$$

and induce her to report the truth.[17]

The principal inherits, in this way, the information of the supervisor and can sign a contract with the agent that induces him to report the truth. This incentive contract is contingent on the information $\hat{\theta}^S$ that the supervisor reports as well as the agent's report $\hat{\theta}^A$. The principal's optimization problem is then subject to the agent's incentive and participation constraints:

$$\begin{aligned} \max_{\{x_{ij}, w_{ij}\}} \quad & \pi \left[q_1 \left(x_{11} - w_{11}\right) + \left(1 - q_1\right)\left(x_{12} - w_{12}\right)\right] \\ & \quad + \left(1 - \pi\right)\left[q_2 \left(x_{21} - w_{21}\right) + \left(1 - q_2\right)\left(x_{22} - w_{22}\right)\right] \\ s.t. \quad & w_{2j} - d(x_{2j} - \theta_2^A) \geq w_{1j} - d(x_{1j} - \theta_2^A) \quad \forall j \in \{1, 2\} \\ & w_{1j} - d(x_{1j} - \theta_1^A) \geq \bar{w} \qquad\qquad\qquad\quad \forall j \in \{1, 2\} \end{aligned} \tag{6}$$

where $x_{ij} = x(\hat{\theta}_i^A, \hat{\theta}_j^S)$ and $w_{ij} = w(\hat{\theta}_i^A, \hat{\theta}_j^S)$ for every $i \in \{1, 2\}$ and every $j \in \{1, 2\}$. Problem (6) is standard. The principal's expected profit is maximized

[16] In this case the optimal mechanism is $[w(\hat{\theta}^A), x(\hat{\theta}^A)]$.

[17] We assume through the analysis that in case of indifference each party behaves in the best possible way for the principal. This tie-breaking rule is used to avoid multiple equilibria in the subgame played by the supervisor and the agent that arises when both parties are indifferent between their actions. The same result could be obtained by augmenting the *honest* mechanism described in this section, as in Ma *et al.* (1988), using nuisance strategies that allow the principal to induce the supervisor and the agent to coordinate on the equilibrium that the principal desires.

subject to the *incentive compatibility* constraints for the high productivity agent and the *individual rationality* constraints for the low productivity agent. We omit individual rationality constraint for the high productivity agent and incentive compatibility constraint for the low productivity agent since, as easily shown, these constraints are not binding in equilibrium.

From Problem (6) we obtain:

$$x_{21} = x_{22} = x_2 > x_{11} > x_{12} \tag{7}$$

$$w_{21} - d(x_2 - \theta_2^A) > w_{22} - d(x_2 - \theta_2^A) > \bar{w} \tag{8}$$

$$w_{11} - d(x_{11} - \theta_1^A) = w_{12} - d(x_{12} - \theta_1^A) = \bar{w} \tag{9}$$

$$w_{21} - d(x_2 - \theta_2^A) = w_{11} - d(x_{11} - \theta_2^A) \tag{10}$$

$$w_{22} - d(x_2 - \theta_2^A) = w_{12} - d(x_{12} - \theta_2^A) \tag{11}$$

These conditions, together with Equation (5), fully characterize what we here label the optimal *honest* contract.

The following Proposition 1 highlights the key feature of the *honest* contract relevant for our analysis of collusion.[18]

Proposition 1. *The premium paid in equilibrium to the high productivity agent is higher if the supervisor reports* $\hat{\theta}_1^S$ *rather than* $\hat{\theta}_2^S$*:* $w_{21} > w_{22}$.

The intuition behind this result is simple. The principal's costs of inducing a high productivity agent to separate himself from a low productivity agent are of two types: a premium, in utility terms, for the high productivity agent and an inefficient effort (output) level that the low productivity agent is required to produce.[19] Whenever the supervisor tells the principal that she thinks the agent has low productivity — that is she has observed a low signal θ_1^S — the principal updates his prior distribution increasing the probability that the agent has a low productivity θ_1^A. This increases, in expected terms, the costs of having the low productivity agent produce an inefficient level of output, while reducing, in expected terms, the costs of a premium for the high productivity agent. Of course, the situation is symmetric and opposite whenever the supervisor reports to the principal a high signal θ_2^S. Therefore, the principal, in equilibrium, trades-off these two costs and offers a higher premium to the high productivity agent, if the supervisor's report is low, than if it is high — Proposition 1 and Equation (8) — and requires the low productivity agent to exert a higher effort, if the supervisor's report is low, than if it is high — Equation (7).

A final question is whether in this world a principal would want to hire a supervisor in the first place. The answer depends on the reservation salary of

[18] For ease of exposition all proofs are presented in the Appendix.

[19] Inefficiency is defined here with respect to an hypothetical first best, obtained in the case the principal observes perfectly the productivity or the effort of the agent.

the supervisor $\bar{s}$. If the constant salary paid to the supervisor does not exceed the principal's gains generated by the availability of the signal θ^S the principal strictly prefers to hire a supervisor.[20]

3 The Possibility of Collusion

3.1 The Collusive Contract

In our setup collusion takes place between two asymmetrically informed parties: the agent and the supervisor. Therefore, in principle it is possible that during the collusion negotiation the uninformed party, the supervisor, learns the private information of the informed party, the agent. Depending on the extensive form of the collusion negotiation this revelation of information might occur before the uninformed party commits to the collusive agreement or after this occurs. The implications of this timing differ depending on how *enforceable* the collusion agreement is: whether the uninformed party can walk away from collusion and what penalties for breach she is required to pay if she does. In modeling collusion we want to specify a general model that encompasses this additional source of information for the supervisor and allows her to walk away from the collusive agreement, possibly at a cost.

The key assumption of our model is that negotiation of the collusive contract takes place between asymmetrically informed parties. This typically leads to multiple equilibria. What matters for our analysis is whether these equilibria separate the two types of agent during the collusion subgame. Indeed, all separating equilibria reveal the type of the agent to the supervisor hence the information of the supervisor improves.

In what follows we do not specify an extensive form for the collusion negotiation game.[21] Instead we follow Laffont and Martimort (1997) and assume the existence of a *collusion designer*. The colluding parties report their private information to the collusion designer. In our setup only the informed party, the agent, reports his private information, we denote this report $\tilde{\theta}_j^A$, $j \in \{1,2\}$. The designer then assigns to the colluding parties an allocation of surplus through the transfer that the agent makes to the supervisor $\beta(\tilde{\theta}_j^A)$, $j \in \{1,2\}$, a given report $\hat{\theta}^S(\tilde{\theta}_j^A)$, $j \in \{1,2\}$, that the supervisor makes to the

[20] The proof that an additional strictly informative signal generates a positive gain to the principal goes as follows. The standard two-tier principal–agent optimization problem can always be written in the form of Problem (6) adding the two constraints $x_{ih} = x_{ik}$, $\forall h \neq k$, $h, k \in \{1,2\}$. These two constraints turn out to be binding at the optimum. Equation (7) shows that this is not true whenever the information reported by the supervisor is available. Thus, the principal is strictly better off in the latter case.

[21] See Felli (1990, Ch. 1) for a closely related model where collusion negotiation proceeds according to a specific extensive form: a take-it-or-leave-it offer from the agent to the supervisor.

principal and a report $\hat{\theta}^A(\tilde{\theta}_j^A)$, $j \in \{1,2\}$ that the agent makes to the principal depending on the agent's report $\tilde{\theta}_j^A$ to the collusion designer. By the revelation principle, without loss of generality we can restrict attention to equilibria of the collusion subgame where the agent reports the truth to the collusion designer, $\tilde{\theta}_j^A = \theta_j^A$: reports are incentive compatible. If equilibrium transfers are such that $\beta(\theta_1^A) \neq \beta(\theta_2^A)$ then the equilibrium of the collusion game is a separating one and the supervisor learns the agent's private information in the collusion subgame. Finally, collusion is a voluntary agreement, hence both parties will agree to participate in the collusion contract only if the allocation induced by the collusive agreement is individually rational. In our environment this implies that the allocation induced by the collusion game has to be strictly better than the allocation induced by the contract offered by the principal if the parties decide not to participate in collusion.[22]

The enforceability of a side contract between two parties is an open issue in the literature on collusion. Often a long-term relationship or a reputational argument is mentioned, in the background, to justify the enforceability of a side contract.[23] In our analysis we use a different approach. We assume that there is an exogenously given penalty for breach, denoted $\kappa(\kappa \geq 0)$, that a party to collusion has to pay to walk out of the collusive agreement. We also assume that a percentage, $\alpha(\alpha \in [0,1])$, of this penalty is received by the counterpart in the collusive agreement. In other words, in the case of breach of the collusive agreement the payoff to the party breaching the side deal decreases by the amount κ while the payoff to the other party of the deal increases by the amount $\alpha\kappa$.

The existing literature has overlooked the principal's ability to prevent collusion by inducing parties to breach their collusive agreement. The difference between the penalty to the party breaching the collusion contract κ and the transfer to the other party $\alpha\kappa$ is meant to capture the fact that most of the cost associated with breaching a side deal is associated with a loss of reputation and future opportunities. This implies that the cost cannot be easily transferred to one's counterpart in collusion. In other words, our general specification encompasses the situation in which *only* the supervisor incurs a loss when breaching the collusive contract ($\alpha = 0$). Alternatively, our setup is also robust to considering situations in which the supervisor can walk away from the collusive contract just before it is signed (that is, the cost for breaching the collusive contract is nil, $\kappa = 0$).

The optimal mechanism we derive below works independently of the size of κ, this is meant to capture the fact that the loss in reputation associated with

[22]Recall our (standard) assumption that when indifferent the agent or the supervisor behaves in the way the principal wants them to behave. In other words, if indifferent neither party participates in collusion.

[23]See Aghion and Caillaud (1988) for a paper that explicitly analyzes this long-term relationship.

the breach may well exceed the financial benefits of the side deal. However, we need the precise value of κ to be common knowledge to all parties. If the principal is not perfectly informed of the nature of the collusive contract and of the penalties associated with its breach, a further level of complexity is added to the corresponding mechanism design problem. We leave the analysis of this setting to further research while noting that in our setting it would be sufficient that the principal knows the upper bound of κ for all of our results to hold true.

3.2 Strongly Collusion-Proof Contracts

As mentioned above, in our setting collusive negotiation takes place between asymmetrically informed parties. This implies that in general the equilibrium outcome of the collusive game is not unique. Hence, the principal's objectives when facing the collusion problem are not at all obvious.

A possible objective might be for the principal to offer a contract to the supervisor and the agent such that when they get involved in the collusion game there exists at least one equilibrium of such a game in which no collusive agreement is enforced. We use a stronger notion of collusion-proofness similar to the one used in the existing literature:

Definition 1. *A contract is strongly collusion-proof if there are no equilibria where a collusive agreement is reached and the equilibrium allocation coincides with the one that arises when both the agent and the supervisor are honest.*

This definition is very restrictive as it requires that there are no equilibria where supervisor and agent collude. Notice that in our setting this restrictive notion comes without loss of generality as we will be able to costlessly avoid collusion.

3.3 Collusion when the Principal Offers the Honest Contract

We begin by observing that the optimal *honest* contract presented in the previous section is not strongly collusion-proof. In other words, if the principal offers such a contract to both the supervisor and the agent, Inequality (8) implies that in the event $\theta^S = \theta_2^S$ the high productivity agent is willing to pay at most

$$b = w_{21} - w_{22} > 0 \tag{12}$$

to the supervisor for her to report $\hat{\theta}_1^S$ while from Equation (9) the low productivity agent is not willing to pay any positive amount to the supervisor for the same report. In the event $\theta^S = \theta_1^S$, instead, neither type of agent is willing to pay any amount to the supervisor for changing her report.

In other words, there exists a whole set of equilibria for the collusion game between the supervisor and the agent in which the supervisor observes the signal

θ_2^S, and the high productivity agent pays a positive bribe to the supervisor to induce him to report a low signal. These equilibria differ depending on the size of the transfer that the high productivity agent pays to the supervisor.

Lemma 1. *The honest contract is not strongly-collusion proof. Under the honest contract, when the supervisor observes the signal θ_2^S, there only exist separating equilibria of the collusion game such that: the high productivity agent pays a positive bribe to the supervisor, $\beta(\theta_2^A) \in (0, b)$, the supervisor reports $\hat{\theta}^S(\theta_2^A) = \hat{\theta}_1^S$ while the agent reports $\hat{\theta}^A(\theta_2^A) = \theta_2^A$. The collusion designer prescribes $\beta(\theta_1^A) = 0$, $\hat{\theta}^S(\theta_1^A) = \hat{\theta}_2^S$ and $\hat{\theta}^A(\theta_1^A) = \theta_1^A$ for the low productivity agent that does not participate in collusion.*

This result shows that the supervisor and the agent may successfully engage in collusion when the *honest* contract is offered to them. A critical feature of Lemma 1 above is that all the equilibria of the collusion game are separating equilibria: $\beta(\theta_2^A) > \beta(\theta_1^A) = 0$. In other words, the high productivity agent reveals his type by participating in collusion and making a positive transfer to the supervisor, whereas the low productivity agent does not. This implies that, in spite of the asymmetry of information that characterizes the collusive negotiation, the supervisor, by observing the agent's willingness to participate in collusion, learns the exact value of the productivity of the agent. In other words, the existence of collusion is synonymous with the agent having high productivity.

4 The Collusion-Proof Optimal Mechanism

In this section we propose a mechanism which allows the principal to prevent collusion between the supervisor and the agent in a costless way: in this mechanism the principal allows the supervisor to blow the whistle by reporting that a collusion agreement has been reached. This report is associated with a premium that the principal pays the supervisor that covers the penalty the supervisor has to pay for breaching the collusive agreement. We now show that such a promise does not involve any extra cost for the principal: it is never carried out in equilibrium. If the high productivity agent observes this clause of the employment contract of the supervisor, he never agrees to participate in collusion as by doing so he loses the informational rent that he otherwise would have gained.

4.1 Candidate Strongly Collusion-Proof Contract

When the *honest* contract is in place, the supervisor may observe two different signals of the agent's productivity: the standard signal $\theta^S \in \Theta^S$ and the information possibly leaked during the collusion game. The latter takes the

form of the agent's truthful report in the collusion game, which, under the *honest* contract, fully reveals the productivity of the agent.

It is critical for the construction of the *collusion-proof* contract to enlarge the message space of the supervisor. This should allow her to report to the principal that the agent has high productivity with certainty once she sees the agent's optimal strategy in the collusion game: $\Theta^S \cup \{\ell_2^A\}$, where ℓ_2^A denotes the leaked information during the collusive game (in equilibrium this will be interpreted as the agent having high productivity). As with every other message, ℓ_2^A is soft or unverifiable information so it is possible that $\ell_2^A \neq \theta_2^A$. The message space of the agent is for the moment left unmodified.

We now specify the part of the strongly collusion-proof candidate contract that concerns the employment contract of the agent. If the supervisor reports any of the messages in Θ^S and the agent reports any of the messages in Θ^A the agent's payoffs are, as in the *honest* contract, characterized by the solution to Problem (6). If the supervisor reports the new message ℓ_2^A, and the agent reports $\hat{\theta}_2^A$ we assume that the agent is asked to produce output x_2, defined in Equation (7), and is paid a wage $\tilde{w}$ so that:

$$\tilde{w} = \bar{w} - \kappa + d(x_2 - \theta_2^A) \tag{13}$$

where $\bar{w}$ is the reservation utility, κ the penalty the supervisor pays to breach the collusive agreement, and $d(x_2 - \theta_2^A)$ is his disutility of effort.

The same remuneration $\tilde{w}$ applies to the agent if the supervisor reports ℓ_2^A and the agent reports $\hat{\theta}_1^A$. Also in this case the agent is asked to produce output x_2.

We now specify the collusion-proof candidate contract between the principal and the supervisor.

If the supervisor reports any of the signals in Θ^S she is paid her constant reservation wage $\bar{s}$, as in Equation (5), whatever the agent's strategy choice.

If the supervisor reports the message ℓ_2^A and the agent reports $\hat{\theta}_2^A$ the supervisor gets her reservation salary plus a premium equal to the penalty κ which she has to pay to breach the collusive agreement and report ℓ_2^A to the principal:

$$s(\ell_2^A, \hat{\theta}_2^A) = \bar{s} + \kappa, \tag{14}$$

Finally, if the supervisor reports ℓ_2^A and the agent reports $\hat{\theta}_1^A$ the supervisor receives her reservation salary minus a positive punishment γ.

$$s(\ell_2^A, \hat{\theta}_1^A) = \bar{s} - \gamma, \quad \gamma > 0 \tag{15}$$

We impose a constraint on the size of the punishment γ so as to prevent the supervisor from reporting ℓ_2^A if the agent does not engage in collusion.

$$p_2\, V(\bar{s} - \gamma) + (1 - p_2)\, V(\bar{s} + \kappa) = V(\bar{s}) \tag{16}$$

Table 1: Collusion-proof contract.

A's report / S's report	$\hat{\theta}_1^A$	$\hat{\theta}_2^A$
ℓ_2^A	$CP^S = [\,\bar{s} - \gamma\,]$ $CP^A = [\,\tilde{w},\, x_2\,]$	$CP^S = [\,\bar{s} + \kappa\,]$ $CP^A = [\,\tilde{w},\, x_2\,]$
$\hat{\theta}_1^S$	$CP^S = [\,\bar{s}\,]$ $CP^A = [\,w_{11},\, x_{11}\,]$	$CP^S = [\,\bar{s}\,]$ $CP^A = [\,w_{21},\, x_{21}\,]$
$\hat{\theta}_2^S$	$CP^S = [\,\bar{s}\,]$ $CP^A = [\,w_{12},\, x_{12}\,]$	$CP^S = [\,\bar{s}\,]$ $CP^A = [\,w_{22},\, x_{22}\,]$

Condition (16) implies that after observing the signal θ_1^S

$$p_1\, V(\bar{s} - \gamma) + (1 - p_1)\, V(\bar{s} + \kappa) < V(\bar{s}) \tag{17}$$

Table 1 summarizes the description of the candidate collusion-proof contract CP.

Finally, recall that if we interpret the supervisor's message ℓ_2^A as informing the principal that collusion occurred then ℓ_2^A is equivalent to the supervisor blowing the whistle on the existence of collusion.

4.2 Preventing Collusion

We can now show that the collusion agreement we presented in Lemma 1 above cannot occur in any equilibrium of the supervisor and agent subgame under the collusion-proof contract.

Proposition 2. *Assume that the principal offers the collusion-proof contract to both the supervisor and the agent and that they both observe the signal* θ_2^S*. There exists no equilibrium collusive agreement such that the supervisor, after observing the signal* θ_2^S*, reports* $\hat{\theta}^S(\theta_2^A) = \hat{\theta}_1^S$ *while the* θ_2^A *agent makes a transfer* $\beta(\theta_2^A) \in (0, b)$ *and reports* $\hat{\theta}^A(\theta_2^A) = \theta_2^A$*.*

The main intuition behind this result is as follows. Recall first that according to Lemma 1 all collusive agreements between the supervisor and the agent under the *honest* contract entail separating equilibria and hence the supervisor learns the true type of the agent.[24] Moreover, the collusion-proof contract specifies payments to both the supervisor and the agent that coincide with the *honest* contract whenever the supervisor reports $\hat{\theta}^S \in \{\theta_1^S, \theta_2^S\}$. However, the collusion-proof contract also offers the supervisor that engaged in collusion the option to breach the collusive agreement at no cost if the supervisor is certain that the agent is high productivity and reports ℓ_2^A to the principal. In the latter case the supervisor payoff is $\bar{s} + \beta(\theta_2^A)$ which coincides with her payoff if she goes along with collusion.[25] The result is that the supervisor is always compensated for breaching the collusion agreement contract. She is thus indifferent and reports ℓ_2^A to the principal. According to the collusion-proof contract the high productivity agent is then strictly better off by not engaging in collusion.

The key observation is that in our framework all equilibria of the collusion game between the supervisor and the agent reveal the exact productivity of the agent to the supervisor. Hence, in equilibrium the supervisor has the discretionary power to report this leaked information to the principal which in turn discourages the agent from participating in collusion. Clearly, the same mechanism would not be successful in preventing pooling collusive agreements if these exist. Costly resources would then be needed to get rid of the pooling equilibria of the collusion game — that is, using a mechanism à la Laffont and Martimort (1997).[26]

[24]There exist only separating equilibria of the collusion game under the *honest* contract because, in equilibrium, the low productivity agent is indifferent whatever the report of the supervisor: the individual rationality constraint of the this type of agent is binding — see Condition (9) and Lemma 1. It follows that the low productivity agent is not willing to pay any amount to change such report. Notice that this is a rather general characteristic of a model where collusion takes place among asymmetrically informed parties.

[25]The supervisor payoff is derived under the assumption that in the collusive contract the bribe $\beta(\theta_2^A)$ is paid upfront and is not refundable. Notice however that $\alpha\kappa$, what the agent receives in the event of a breach of the collusive contract on the part of the supervisor, may well exceed the bribe paid upfront and hence be regarded as a refund of this bribe in the event of a breach.

[26]Notice that similar considerations also apply when the supervisor has *hard* information as in Tirole (1986). Indeed, in this case collusion only occurs between symmetrically informed parties, hence costly resources need to be used to prevent collusion arising in equilibrium.

Notice that the existence of any level of asymmetric information is enough for whistle-blowing to be successful in eliminating "some" collusion. Indeed, CP prevents collusive agreement from arising in equilibrium whatever the precision of the signal θ_2^S with the exception of the limit case $q_1 = 1$, that is when the supervisor is perfectly informed after observing a high signal. The following result identifies the situations in which our result on collusion-proofness holds.

Corollary 1. *No equilibrium collusive agreement, as in Lemma 1, exists for any imperfect signal θ_2^S observed by the supervisor:* $(1 - p_2) < 1$.

This result implies that the costs of preventing collusion are discontinuous in the limit as the noise associated with the signal observed by the supervisor vanishes. When the supervisor's signal is perfect, preventing collusion becomes costly for the principal as in Laffont and Martimort (1997). The supervisor and the agent collude only when the supervisor perfectly observes the productivity of the agent. Allowing the supervisor to blow the whistle is then of no use to the principal.

It is worth observing that whistle-blowing allows the principal to prevent collusion in a costless way when the signal θ_1^S is a perfect signal: when q_2 converges to zero, or equivalently p_1 converges to one. Collusion occurs only when both the supervisor and the agent observe the signal θ_2^S, which is not a perfect signal: the probability that the agent has low productivity is not null, $p_2 > 0$.[27] This implies that the collusion game can still reveal some information to the supervisor. Therefore allowing the supervisor to blow the whistle is still effective in preventing collusion agreements between the supervisor and the agent at no additional costs for the principal (Proposition 2 applies).

A final observation concerns the willingness of the principal to use a supervisor. Since the solution to the collusion problem we propose is costless for the principal, the same considerations we presented above apply in this case. There exist values of the reservation salary of the supervisor for which the principal has a strictly positive gain by hiring her.

[27] Notice that if θ_1^S is a perfect signal, collusion between the supervisor and the agent will take a slightly different form even when the *honest* contract is offered. In fact, in such a case there will be no need for the principal to specify a payoff for the high productivity agent if the supervisor reports $\hat{\theta}_1^S$: the signal is perfect, so, provided the supervisor reports the truth, the agent's productivity is certainly low. However, it is still profitable for the high productivity agent to bribe the supervisor to report signal $\hat{\theta}_1^S$ when θ_2^S is observed but this report requires the agent to report $\hat{\theta}_1^A$ or, equivalently, to produce a low output. Indeed, the premium the high productivity agent receives in this way is greater than the informational rent he would get if the supervisor reports the truth.

5 Blackmail

5.1 The Blackmail Threat

A different type of manipulation may occur when the principal offers the collusion-proof contract. This takes the form of *blackmail*. Blackmail is a unilateral threat by one of the parties involved. For blackmail to occur in equilibrium we require the threat to be credible. In other words, an equilibrium of our model with blackmail is a PBE where the blackmailing party asks for a transfer $\mu > 0$ from the blackmailed party. If the transfer is made then the blackmailing party complies with the equilibrium strategy in the absence of blackmail. If the blackmailed party refuses to make the transfer the blackmailing party behaves consistently with his/her threat.

In what follows we show that allowing the supervisor to report the information leaked during the collusion game to the principal provides the supervisor with the opportunity to abuse his authority and threaten the agent to report ℓ_2^A, even in the absence of any collusion, unless she receives a payment μ on the part of the agent.

Assume that the principal offers the collusion-proof contract to both the supervisor and the agent and they both accept. As from Proposition 2, whatever the signal θ_i^S, the agent refuses to participate in any collusive agreement with the supervisor. Assume now that, before reporting its observed signal, the supervisor threatens the agent to report the signal ℓ_2^A unless she receives a strictly positive payment $\mu > 0$. We show below that there exist values of μ that render this blackmail a credible threat on the part of the supervisor.

Notice first that since no collusion took place, following this threat, if both types of agent behave in the same way, the supervisor is still uninformed on the value of the agent's type θ_i^A. Therefore, Equations (16) and (17) guarantee that under contract CP if both types of agent refuse or accept to pay μ the supervisor reports θ_i^S rather than ℓ_2^A. In other words the blackmail threat is not credible.

This means that blackmail can be credible only if the two types of agent make different choices when deciding whether or not to succumb to blackmail. We, then, need to consider each type of agent's willingness to go along with the supervisor's blackmail threat.

Consider first the θ_2^A agent. His maximum willingness to pay to avoid the supervisor reporting the message ℓ_2^A is given by the difference between his payoff if the supervisor reports the signal θ_i^S and his payoff when she reports ℓ_2^A. This difference is

$$\mathcal{W}_2^A = \kappa + d(x_{1i} - \theta_1^A) - d(x_{1i} - \theta_2^A) \tag{18}$$

Consider now the θ_1^A agent. His maximum willingness to pay is instead

$$\mathcal{W}_1^A = \kappa + d(x_2 - \theta_1^A) - d(x_2 - \theta_2^A) \tag{19}$$

It then follows from the convexity of the disutility of effort, $d''(\cdot) > 0$, and $x_2 > x_{1i}$ that the θ_1^A agent is willing to pay strictly more than the θ_2^A agent to avoid the supervisor reporting ℓ_2^A rather than θ_i^S.

$$\mathcal{W}_1^A > \mathcal{W}_2^A \tag{20}$$

This implies that the only blackmail payment μ that separates the two types of agent is one that is acceptable to the θ_1^A agent but not acceptable to a θ_2^A agent:

$$\mathcal{W}_1^A > \mu > \mathcal{W}_2^A \tag{21}$$

Any other value of μ is either acceptable to both types of agent or not acceptable to either type.

Notice now that if the supervisor blackmails the agent by asking for a transfer μ satisfying Inequality (21), the supervisor following the agent decision to pay or not μ discovers the type of the agent. In particular, following the agent's decision not to pay μ, the supervisor knows that the agent is of type θ_2^A. It is then optimal for the supervisor to report ℓ_2^A since in so doing she gets payoff $\bar{s} + \kappa$ rather than the payoff $\bar{s}$ she gets from reporting θ_i^S. On the other hand, following the agent's decision to pay μ the supervisor knows that the agent is of type θ_1^A. It is then optimal for the superior to report the signal θ_i^S since in so doing she gets from the principal $\bar{s}$ instead of $\bar{s} - \gamma$ as from the collusion-proof contract. In other words, the supervisor's blackmail threat is credible.

Lemma 2. *The contract CP is vulnerable to blackmail. Under contract CP there exists a set of equilibria such that if both types of agent and the supervisor accept the collusion-proof contract the supervisor credibly threatens the agent to report ℓ_2^A unless a transfer $\mu > 0$ satisfying Inequality (21), is paid by the agent to the supervisor.*

As in the case of collusion, the reason why the supervisor's blackmail is credible is that if blackmail occurs it separates the two types of agents and, as a consequence, the supervisor discovers the type of the agent and finds it optimal to behave differently depending on whether the agent accepts or not to pay the transfer μ.

5.2 Preventing Collusion and Blackmail

We propose here a mechanism that prevents both collusion and blackmail. Similar to the collusion-proof contract, this mechanism now also allows the agent to blow the whistle by enlarging his message space and allowing him to report whether he received a blackmail threat from the supervisor. We denote the latter message $\mathcal{B}$. We first define what it means for a contract to be *blackmail-proof*.

Definition 2. *A contract is blackmail-proof if all PBE of the model are such that no party can credibly threaten any other party to deviate from the prescribed equilibrium behavior in the absence of blackmail in exchange for a transfer* $\mu > 0$.[28]

We propose a new contract, *collusion and blackmail-proof* (denoted CBP), that allows the principal to prevent, costlessly, both collusion and blackmail. This contract enlarges the message space of both the supervisor and the agent by allowing both to blow the whistle. It not only allows the supervisor to report the message ℓ_2^A if collusion occurs, but also allows the agent to report to the principal that blackmail occurred, message $\mathcal{B}$, and in so doing triggers a penalty for the supervisor that discourages her from blackmailing the agent.

Following the supervisor's report ℓ_2^A, we allow the agent to report either his type $\hat{\theta}_i^A$ or the additional message $\mathcal{B}$. Notice that we restrict the agent to report the additional message $\mathcal{B}$ only after the supervisor's report of the leaked information ℓ_2^A in order to prevent additional opportunities for manipulation of the optimal mechanism on the part of both the supervisor and the agent.

In case the agent reports $\mathcal{B}$ the supervisor's remuneration is $\bar{s} - \varepsilon$, where $\varepsilon > 0$ is an arbitrary small number, while the agent is asked to produce output x_2, defined in Equation (7), and is paid a wage $\tilde{w}$ defined in Equation (13). In other words, the agent is indifferent between reporting any of the messages in $(\hat{\theta}_1^A, \hat{\theta}_2^A)$ and $\mathcal{B}$ while the supervisor, if she does not know the type of the agent, is strictly worse off when the agent reports $\mathcal{B}$. We summarize the contract CBP in Table 2.

We are now in a position to show that if the principal offers contract CBP to both the supervisor and the agent and this contract is accepted, no collusion or blackmail takes place and the outcome coincides, on the equilibrium path, with the *honest* contract above.

Proposition 3. *The contract* CBP *is strongly collusion-proof and it is not liable for any form of blackmail either on the part of the superior or the agent. The PBE of the continuation game between the supervisor and the agent coincides with the PBE of the corresponding continuation game under the honest contract (Proposition 1).*

Intuitively, allowing both the Supervisor and the Agent to blow the whistle makes the supervisor strictly worse off by blackmailing the agent (her payoff is $\bar{s} - \varepsilon$) rather than not engaging in any threat (her payoff is $\bar{s}$) because the agent is indifferent between any report and will thus report $\mathcal{B}$. In the Appendix we also show that the message $\mathcal{B}$ does not introduce new opportunities for both

[28] Recall that in our model we assume that when indifferent, a party complies with the strategy desired by the principal, hence blackmail is not credible if, in the subgame following the blackmailed party decision to not pay μ, the blackmailing party is indifferent to complying or not with the blackmail threat.

Table 2: Collusion and blackmail-proof contract (CBP).

A's report / S's report	$\hat{\theta}_1^A$	$\hat{\theta}_2^A$	$\mathcal{B}$
ℓ_2^A	$CBP^S = [\,\bar{s} - \gamma\,]$ $CBP^A = [\,\tilde{w},\, x_2\,]$	$CBP^S = [\,\bar{s} + \kappa\,]$ $CBP^A = [\,\tilde{w},\, x_2\,]$	$CBP^S = [\,\bar{s} - \varepsilon\,]$ $CBP^A = [\,\tilde{w},\, x_2\,]$
$\hat{\theta}_1^S$	$CBP^S = [\,\bar{s}\,]$ $CBP^A = [\,w_{11},\, x_{11}\,]$	$CBP^S = [\,\bar{s}\,]$ $CBP^A = [\,w_{21},\, x_{21}\,]$	
$\hat{\theta}_2^S$	$CBP^S = [\,\bar{s}\,]$ $CBP^A = [\,w_{12},\, x_{12}\,]$	$CBP^S = [\,\bar{s}\,]$ $CBP^A = [\,w_{22},\, x_{22}\,]$	

the supervisor and the agent to engage in collusion. It also prevents the agent from blackmailing the supervisor with the additional message $\mathcal{B}$.

6 Concluding Remarks

In this paper we have showed that potential collusion between the intermediate and bottom layers of a hierarchy might make it desirable for the principal to allow his subordinates to blow the whistle. In other words, the principal, allowing the supervisor to blow the whistle on whether collusion occurred, aligns the objectives of the supervisor with her own and eliminates the opportunity of collusive behavior between layers and, ultimately, enhances efficiency. However, allowing the supervisor to blow the whistle introduces the opportunity for the supervisor to blackmail the agent. Once again, allowing this time the agent to blow the whistle on the existence of blackmail prevents any wrongdoing and replicates the honest outcome.

At the core of our argument is the idea that in many collusive agreements or blackmail interactions some information is shared between collusive parties and this information can subsequently be used to the detriment of the collusive parties or the blackmailer. In such circumstances the principal (residual claimant

in our setting) could appropriately reward the leaking of such information and in this manner avoid collusive agreements and blackmail threats.

In principle, one could apply this same logic to the contract the principal writes with the agent. Indeed, the direct mechanism we have analyzed specifies that the agent and supervisor report their private information to the principal and only at a later stage can the agent exert her productive effort. In principle, one could argue that the principal could use the information revealed by the agent and renege on the promised output-contingent remuneration before output is realized. Notice however, that this multistage performance on the part of the agent only applies to the direct revelation mechanism. The most obvious indirect mechanism would have the agent, after the supervisor's report, exert a productive effort that leads to the observed output. In other words, the agent's private information is only revealed when production is completed. Assuming that the principal can renege on a promise at this stage would be equivalent to assuming that simple trade cannot be enforced. This lack of commitment clearly opens up further sources of inefficiency in usual contracting environments that are beyond the scope of this paper. We should add that it is sensible to assume that contracts written by the principal can be enforced by a Court of Law while the same cannot be said of side contracts. Our paper should then be read as a first step in the direction of explicitly allowing contracting parties to breach their signed agreements.

Finally, our analysis sheds light on the use of rules versus discretion in the designing of the optimal degree of decision power of members of an organization. Tirole (1986) argues that fixed rules as opposed to discretion might be used to reduce patterns of collusive behavior in large (private or public) organizations. In this paper we show that whenever collusion takes place in conditions of asymmetric information, an increase in the discretionary power (the message space) of intermediate layers of the organization as emphasised in McCubbins and Schwartz (1984), may have a beneficial effect in reducing the possibility of both collusion and blackmail.

The results presented in this paper can be interpreted as a way to implement a particular outcome that enlarges the strategy space of subordinates. The basic intuition is as follows.[29] We have learned from the literature on commitment that under certain conditions a player can increase his welfare whilst restricting, in a credible way, his choices: his strategy space.[30] This paper complements this literature by showing that enlarging the strategy space of members of an organization (i.e. delegating oversight activities to them) enhances the welfare of the residual claimant of an activity whilst reducing the welfare of his subordinate.

[29] We are indebted to David Canning for this intuition.

[30] See for example Laffont and Tirole (1988).

Appendix

Proof of Lemma 1:
Assume that the *honest* contract binds the principal, the supervisor and the agent. Assume that S observes the signal θ_2^S. The collusion contracts $\mathcal{C} = \{\beta(\theta_1^A), \beta(\theta_2^A); \hat{\theta}^S(\theta_2^A) = \hat{\theta}_1^S, \hat{\theta}^S(\theta_2^A) = \hat{\theta}_2^S\}$, such that $\beta(\theta_1^A) = 0$, $\beta(\theta_2^A) \in (0, b)$ — with b as in Equation (12) — and S reports $\hat{\theta}_1^S$ if the agent reports θ_2^A to the collusion designer and $\hat{\theta}_2^S$ if the agent reports θ_1^A to the collusion designer, satisfies the collusion-game incentive compatibility and individual rationality constraints for the θ_2^A agent but does not satisfy the collusion-game individual rationality constraint for the θ_1^A agent.

Consider first the collusion game incentive compatibility constraint for the θ_2^A agent

$$w_{21} - \beta(\theta_2^A) - d(x_2 - \theta_2^A) \geq w_{12} - \beta(\theta_1^A) - d(x_{12} - \theta_2^A)$$
$$= w_{22} - \beta(\theta_1^A) - d(x_2 - \theta_2^A) \tag{A.1}$$

Equations (8) and (12), together with $0 < \beta(\theta_2^A) < b$ and $\beta(\theta_1^A) = 0$, imply that Condition (A.1) holds with a strict inequality.[31] Consider now the collusion-game individual rationality constraint for the θ_2^A agent:

$$w_{21} - \beta(\theta_2^A) - d(x_2 - \theta_2^A) \geq w_{22} - d(x_2 - \theta_2^A) \tag{A.2}$$

Equations (8), (12) together with $0 < \beta(\theta_2^A) < b$ imply that Condition (A.2) also holds with a strict inequality. The collusion-game individual rationality constraint for the θ_1^A agent is instead:

$$w_{12} - \beta(\theta_1^A) - d(x_{12} - \theta_1^A) \geq w_{12} - d(x_{12} - \theta_1^A) \tag{A.3}$$

From $\beta(\theta_1^A) = 0$ it follows that Condition (A.3) holds with equality. This means that the θ_1^A agent does not participate in the collusion game since, when indifferent, the agent does what the principal would like him to do.[32]

Consider now the supervisor's collusion-game individual rationality constraint associated with the collusion contract $\mathcal{C}$. This is:

$$\nu V(\bar{s}) + (1 - \nu)V(\bar{s} + \beta(\theta_2^A)) \geq V(\bar{s}) \tag{A.4}$$

where ν denotes the supervisor's beliefs at the collusion stage that the type of the agent is θ_1^A. Clearly, if $\nu < 1$ and $\beta(\theta_2^A) > 0$ Constraint (A.4) is satisfied with a strict inequality. In other words, under the *honest* contract it is an equilibrium of the collusion game for both the type θ_2^A agent and the supervisor to accept any of the collusion contracts $\mathcal{C}$.

[31] Notice that Equation (11) and $\beta(\theta_1^A) = 0$ imply that, following the deviation of the θ_2^A agent in the report to the collusion designer, this agent will be indifferent when making his report to the principal and hence will report the truth.

[32] Notice that a similar argument shows that neither the θ_1^A agent nor the θ_2^A participate in collusion if the contract $\mathcal{C}$ is such that $\beta(\theta_2^A) = b$.

Proof of Proposition 2:
Assume that the supervisor observes the signal θ_2^S. Consider any incentive compatible collusive contract such that the supervisor reports $\theta^S(\theta_2^A) = \hat{\theta}_1^S$ and the θ_2^A agent pays the bribe $\beta(\theta_2^A) \in (0, b)$, as in Equation (12). We proceed in four steps.

Step 1. *The agent always reports the truth to the principal whatever his productivity and the outcome of the collusion game.*

We start from the high productivity agent. Assume that the θ_2^A agent participates in collusion and the collusion contract is not breached by the supervisor this agent's payoff is then either $w_{21} - \beta(\theta_2^A) - d(x_2 - \theta_2^A)$, if he reports the truth, or $w_{11} - \beta(\theta_2^A) - d(x_{11} - \theta_2^A)$, if he does not. Equation (10) implies that the agent is indifferent between these two payoffs, hence he reports the truth. Assume next that the θ_2^A agent participates in collusion and the collusion contract is breached by the supervisor who reports the message ℓ_2^A. This agent's payoff is then $\tilde{w} - \beta(\theta_2^A) + \alpha\kappa - d(x_2 - \theta_2^A) = \bar{w} - \beta(\theta_2^A) - (1-\alpha)\kappa$ whether he reports the truth or he does not. Hence, the agent, being indifferent, reports the truth.

Assume now that the θ_2^A agent does not participate in collusion. If the supervisor reports the observed signal θ_2^S, the binding incentive compatibility constraint for the high productivity agent, Equation (11), implies that the agent reports the truth. If instead the supervisor reports the additional signal ℓ_2^A the agent's payoff is $\tilde{w} - d(x_2 - \theta_2^A) = \bar{w} - \kappa$ whether he reports the truth or he does not. Hence, the high productivity agent is indifferent and reports the truth.

Consider, now, the low productivity agent. Assume that the θ_1^A agent participates in collusion and the supervisor does not breach the collusive agreement and reports the signal $\hat{\theta}_1^S$, the agent's payoff is $\bar{w} - \beta(\theta_1^A)$ if he reports $\hat{\theta}_1^A$ and $w_{21} - \beta(\theta_1^A) - d(x_2 - \theta_1^A) < \bar{w} - \beta(\theta_1^A)$ if he reports $\hat{\theta}_2^A$. Hence the agent will report the truth. Conversely, assume that the θ_1^A agent does not participate in collusion and the supervisor reports the observed signal $\hat{\theta}_2^S$. The incentive compatibility constraint for the low productivity agent — implied by $\theta_2^A > \theta_1^A$ and Equation (11) — holds with strict inequality, hence the agent reports the truth.

Finally, assume the supervisor reports the message ℓ_2^A, the agent's payoff is the same whatever his report. It is either $\tilde{w} - \beta(\theta_1^A) + \alpha\kappa - d(x_2 - \theta_1^A) = \bar{w} - \beta(\theta_1^A) - (1-\alpha)\kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$, if the supervisor participated in collusion, breaches the collusion contract and reports the message ℓ_2^A or is $\tilde{w} - d(x_2 - \theta_1^A) = \bar{w} - \kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$ if she reports the message ℓ_2^A without participating in collusion. Either case, the agent, being indifferent, reports the truth.

Step 2. *Derivation of the supervisor's best response when the supervisor observes* θ_2^S.

As above, denote ν the supervisor's belief that the agent is of type θ_1^A at the collusion stage. Assume, first, that both the agent and the supervisor accept to participate in collusion. The supervisor's payoff is then $V(\bar{s}+\beta(\theta_2^A))$ if she complies with the collusion contract and reports $\theta^S(\theta_2^A) = \hat{\theta}_1^S$. The supervisor expected payoff is instead $\nu V(\bar{s}+\beta(\theta_2^A)-\gamma-\kappa)+(1-\nu)V(\bar{s}+\beta(\theta_2^A))$ if she breaches the collusion contract and reports the additional signal ℓ_2^A. Equation (14) implies that if $\nu > 0$ the former option yields a higher payoff to the supervisor, hence she will comply with the collusion contract. If, instead, $\nu = 0$ the supervisor is indifferent between the two options, hence she acts in the way most preferred by the principal: she breaches the collusion contract and reports the signal ℓ_2^A to the principal.

Consider now the supervisor decision whether to participate in the collusion game. As seen above her payoff, whether she breaches the collusion contract or not, is $V(\bar{s}+\beta(\theta_2^A))$ while her payoff is $V(\bar{s})$ if she refuses to participate in the collusion game and reports the observed signal $\hat{\theta}_2^S$. Clearly, if $\beta(\theta_2^A) > 0$ the supervisor is better off accepting to participate in collusion. Only if $\beta(\theta_2^A) = 0$ the supervisor is indifferent and refuses to participate in the collusion game.

Step 3. *The value of the supervisor's belief* $\nu = 0$ *is the only one consistent with the low productivity agent's behavior.*

Assume $\nu > 0$ and consider the behavior of the low productivity agent. Given the supervisor's best response, Step 2, the agent's payoff is either $\bar{w} - \beta(\theta_1^A)$, if he participates in the collusion game, or $\bar{w}$, if he does not. Clearly the low productivity agent always refuses to participate in the collusion game for $\beta(\theta_1^A) \geq 0$. This contradicts the hypothesis $\nu > 0$.

Step 4. *The agent always refuses to participate in the collusion game whatever his productivity.*

We start from the low productivity agent. Given Steps 2 and 3, the agent's payoff is $\tilde{w} + \alpha\kappa - d(x_2 - \theta_1^A) = \bar{w} - (1-\alpha)\kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$ if he participates in the collusion game and produces output x_2. Such payoff is strictly lower than the agent's reservation wage $\bar{w}$ since $\theta_1^A < \theta_2^A$. Conversely, if the agent refuses to participate in the collusion game his payoff is $\bar{w}$, by Equation (9). Hence, the low productivity agent refuses to participate in the collusion game.

Finally, consider the high productivity agent. Steps 1, 2 and 3 imply that if he accepts the collusive offer his payoff is $\tilde{w} - \beta(\theta_2^A) + \alpha\kappa - d(x_2 - \theta_2^A) = \bar{w} - \beta(\theta_2^A) - (1-\alpha)\kappa$. Conversely, if he rejects the collusive offer his payoff is $w_{22} - d(x_{22} - \theta_2^A)$ which, by Equation (8), is strictly greater than

$\bar{w} - \beta(\theta_2^A) - (1-\alpha)\kappa$. Hence, the high productivity agent refuses to participate in the collusion game.

Proof of Corollary 1:
The proof follows immediately from the observation that the proof of Proposition 2 does not rely on how precise the signal θ_2^S is of the agent being of type θ_2^A but only on whether, having observed signal θ_2^S there still exists a residual, strictly positive, probability that the agent is of type θ_1^A.

Proposition 3 will be proved with the assistance of the following four lemmas:

Lemma 3. *The contract CBP is such that, in the absence of collusion, any blackmail threat on the part of S is not credible.*

Proof. Assume S threatens A to report ℓ_2^A unless she receives a transfer $\mu > 0$ from A. Notice, first, that following S's report ℓ_2^A, A is indifferent and hence by assumption he reports $\mathcal{B}$ to P.

Assume now that while the type θ_1^A of A pays $\mu > 0$ type θ_2^A of A does not. Following A's decision not to pay, S's payoff is $V(\bar{s} - \varepsilon)$ if S reports ℓ_2^A and $V(\bar{s})$ if S reports θ_i^S. Therefore, S reports θ_i^S and hence S's threat is not credible.

Assume now that while the type θ_1^A of A does not pay $\mu > 0$ type θ_2^A of A does. Following A's decision not to pay, S's payoff is $V(\bar{s} - \varepsilon)$ if S reports ℓ_2^A and $V(\bar{s})$ if S reports θ_i^S. Once again, S reports θ_i^S and hence S's threat is not credible.

Assume now that both types of A pay $\mu > 0$. Following type θ_2^A of A's deviation not to pay, S's payoff is $V(\bar{s} - \varepsilon)$ if S reports ℓ_2^A and $V(\bar{s})$ if S reports θ_i^S whatever S's beliefs. Once again, S reports θ_i^S and hence S's threat is not credible.

Finally, consider the case where both types of A do not pay $\mu > 0$. S's payoff is $V(\bar{s} - \varepsilon)$ if S reports ℓ_2^A and $V(\bar{s})$ if S reports θ_i^S. In other words, S reports θ_i^S and S's threat is not credible.

Lemma 4. *The contract CBP is such that, in the absence of collusion, any blackmail threat on the part of A is not credible.*

Proof. Assume that, in the absence of any threat on the part of S, A threatens S to report $\mathcal{B}$ following S's report ℓ_2^A unless he receives a transfer $\eta > 0$ from S.

Consider first the case in which the type θ_1^A of A threatens S while the type θ_2^A of A does not. Following A's threat, S updates her beliefs that A is of type θ_1^A. S's payoff if she reports θ_i^S is $V(\bar{s})$ if S does not pay and $V(\bar{s} - \eta)$ if she pays. S's payoff if she reports ℓ_2^A depends on A's report. Notice, however,

that A's payoff is the same whatever his report. In particular, A's payoff is $\tilde{w} - d(x_2 - \theta_1^A) = \bar{w} - \kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$ if S does not pay and $\tilde{w} - d(x_2 - \theta_1^A) + \eta = \bar{w} - \kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A) + \eta$ if S pays. In either case, A is indifferent among his messages and hence reports the truth θ_1^A. Given this report, S's payoff is then $V(\bar{s} - \gamma)$ if she does not pay and $V(\bar{s} - \gamma - \eta)$ if she pays. This implies that S's optimal strategy is not to pay η and to report θ_i^S while A's optimal strategy is to report θ_1^A. In other words, A's threat is not credible.

Consider next the case in which the type θ_1^A of A does not threaten S while the type θ_2^A of A does. Following A's threat, S updates her beliefs that A is of type θ_2^A. S's payoff if she reports θ_i^S is $V(\bar{s})$ if S does not pay and $V(\bar{s} - \eta)$ if she pays. Once again S's payoff if she reports ℓ_2^A depends on A's report. Notice, however, that as above A's payoff is the same whatever his report. In particular, A's payoff is $\tilde{w} - d(x_2 - \theta_2^A) = \bar{w} - \kappa$ if S does not pay and $\tilde{w} - d(x_2 - \theta_2^A) + \eta = \bar{w} - \kappa + \eta$ if S pays. In either case, A is indifferent among his messages and hence reports the truth θ_2^A. Given this report S's payoff is then $V(\bar{s} + \kappa)$ if she does not pay and $V(\bar{s} + \kappa - \eta)$ if she pays. This implies that S's optimal strategy is not to pay η and to report ℓ_2^A while A's optimal strategy is to report θ_2^A. Once again, A's threat is not credible.

Finally, consider the case in which both types of A threaten S. Following A's threat, S does not update her beliefs. Now S's payoff if she reports θ_i^S is $V(\bar{s})$ if S does not pay and $V(\bar{s} - \eta)$ if she pays. Once again S's payoff if she reports ℓ_2^A depends on A's report. Notice, however, that as above A's payoff is the same whatever his report. In particular, type θ_j^A, $j \in \{1, 2\}$, of A's payoff is $\tilde{w} - d(x_2 - \theta_j^A) = \bar{w} - \kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_j^A)$ if S does not pay and $\tilde{w} - d(x_2 - \theta_j^A) + \eta = \bar{w} - \kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_j^A) + \eta$ if S pays. In either case, A is indifferent among his messages and hence reports the truth θ_j^A. Given this report by Conditions (16) and (17) S's payoff is smaller than $V(\bar{s})$ if she does not pay and smaller than $V(\bar{s} - \eta)$ if she pays. This implies that S's optimal strategy is not to pay η and to report θ_i^S while A's optimal strategy is to report θ_j^A. Once again, A's threat is not credible.

Lemma 5. *The contract CBP is such that, if the supervisor observes the signal θ_2^S, there exists no equilibrium collusive agreement such that the supervisor, after observing the signal θ_2^S, reports the message $\theta^S(\theta_2^A) = \hat{\theta}_1^S$ and the θ_2^A agent makes a transfer $\beta(\theta_2^A) \in (0, b)$, as from Equation (12).*

Proof. Assume the supervisor observes the signal θ_2^S. We proceed in four steps.

Step 1. *The agent always reports the truth to the principal whatever his productivity and the outcome of the collusion game.*

We start from the high productivity agent. Assume that the θ_2^A agent participates in collusion and the collusion contract is not breached by the

supervisor this agent's payoff is then either $w_{21} - \beta(\theta_2^A) - d(x_2 - \theta_2^A)$, if he reports the truth, or $w_{11} - \beta(\theta_2^A) - d(x_{11} - \theta_2^A)$, if he does not. Equation (10) implies that the agent is indifferent between these two payoffs, hence he reports the truth. Assume next that the θ_2^A agent participates in collusion and the collusion contract is breached by the supervisor who reports the message ℓ_2^A. This agent's payoff is then $\bar{w} - \beta(\theta_2^A) - (1-\alpha)\kappa$ whether he reports $\hat{\theta}_2^A$, $\hat{\theta}_2^A$ or $\mathcal{B}$. Hence, the agent, being indifferent, reports the truth.

Assume now that the θ_2^A agent does not participate in collusion. If the supervisor reports the observed signal θ_2^S, the binding incentive compatibility constraint for the high productivity agent, Equation (11), implies that the agent reports the truth. If instead the supervisor reports the additional signal ℓ_2^A the agent's payoff is $\bar{w} - \kappa$ whether he reports $\hat{\theta}_2^A$, $\hat{\theta}_2^A$ or $\mathcal{B}$. Once again, the high productivity agent being indifferent reports the truth.

Consider, now, the low productivity agent. Assume that the θ_1^A agent participates in collusion and the supervisor does not breach the collusive agreement and reports the signal $\hat{\theta}_1^S$, the agent's payoff is $\bar{w} - \beta(\theta_1^A)$ if he reports $\hat{\theta}_1^A$ and $w_{21} - \beta(\theta_1^A) - d(x_2 - \theta_1^A) < \bar{w} - \beta(\theta_1^A)$ if he reports $\hat{\theta}_2^A$. Hence the agent will report the truth. Conversely, assume that the θ_1^A agent does not participate in collusion and the supervisor reports the observed signal $\hat{\theta}_2^S$. The incentive compatibility constraint for the low productivity agent — implied by $\theta_2^A > \theta_1^A$ and Equation (11) — holds with strict inequality, hence the agent reports the truth. Finally, assume the supervisor reports the additional signal ℓ_2^A, the agent's payoff is the same whether he reports $\hat{\theta}_2^A$, $\hat{\theta}_2^A$ or $\mathcal{B}$. It is either $\tilde{w} - \beta(\theta_1^A) + \alpha\kappa - d(x_2 - \theta_1^A) = \bar{w} - \beta(\theta_1^A) - (1-\alpha)\kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$, if the supervisor participated in collusion, breaches the collusion contract and reports the message ℓ_2^A or is $\tilde{w} - d(x_2 - \theta_1^A) = \bar{w} - \kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$ if she reports the message ℓ_2^A without participating in collusion. Either case, the agent, being indifferent, reports the truth.

Step 2. *Derivation of the supervisor's best response when the supervisor observes θ_2^S and does not blackmail the agent.*

Once again, denote ν the supervisor's belief that the agent is of type θ_1^A at the collusion stage. Assume, first, that both the agent and the supervisor accept to participate in collusion. The supervisor's payoff is then $V(\bar{s} + \beta(\theta_2^A))$ if she complies with the collusion contract and reports $\theta^S(\theta_2^A) = \hat{\theta}_1^S$. The supervisor expected payoff is instead $\nu V(\bar{s} + \beta(\theta_2^A) - \gamma - \kappa) + (1-\nu)V(\bar{s} + \beta(\theta_2^A))$ if she breaches the collusion contract and reports the additional signal ℓ_2^A. Equation (14) implies that if $\nu > 0$ the former option yields a higher payoff to the supervisor, hence she will comply with the collusion contract. If, instead, $\nu = 0$ the supervisor is indifferent between the two options, hence she acts in the way most preferred by the principal: she breaches the collusion contract and reports the signal ℓ_2^A to the principal.

Consider now the supervisor decision whether to participate in the collusion game. As seen above her payoff, whether she breaches the collusion contract or not, is $V(\bar{s}+\beta(\theta_2^A))$ while her payoff is $V(\bar{s})$ if she refuses to participate in the collusion game and reports the observed signal $\hat{\theta}_2^S$. Clearly, if $\beta(\theta_2^A) > 0$ the supervisor is better off accepting to participate in collusion. Only if $\beta(\theta_2^A) = 0$ the supervisor is indifferent and refuses to participate in the collusion game.

Step 3. *If collusion occurs the value of the supervisor's belief $\nu = 0$ is the only one consistent with the low productivity agent's behavior.*

Assume $\nu > 0$ and consider the behavior of the low productivity agent. Given the supervisor's best response, Step 2, the agent's payoff is either $\bar{w} - \beta(\theta_1^A)$, if he participates in the collusion game, or $\bar{w}$, if he does not. Clearly the low productivity agent always refuses to participate in the collusive for $\beta(\theta_1^A) \geq 0$. This contradicts the hypothesis $\nu > 0$.

Step 4. *The agent always refuses to participate in the collusion game whatever his productivity.*

We start from the low productivity agent. Given Steps 2 and 3, the agent's payoff is $\tilde{w} + \alpha\kappa - d(x_2 - \theta_1^A) = \bar{w} - (1-\alpha)\kappa + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A)$ if he participates in the collusion game and produces output x_2. Such payoff is strictly lower than the agent's reservation wage $\bar{w}$ since $\theta_1^A < \theta_2^A$. Conversely, if the agent refuses to participate in the collusion game his payoff is $\bar{w}$, by Equation (9). Hence, the low productivity agent refuses to participate in the collusion game.

Finally, consider the high productivity agent. Steps 1, 2 and 3 imply that if he accepts the collusive offer his payoff is $\tilde{w} - \beta(\theta_2^A) + \alpha\kappa - d(x_2 - \theta_2^A) = \bar{w} - \beta(\theta_2^A) - (1-\alpha)\kappa$. Conversely, if he rejects the collusive offer his payoff is $w_{22} - d(x_{22} - \theta_2^A)$ which, by Equation (8), is strictly greater than $\bar{w} - \beta(\theta_2^A) - (1-\alpha)\kappa$. Hence, the high productivity agent refuses to participate in the collusion game.

Lemma 6. *The contract CBP is such that there exists no equilibrium collusive agreement where:*

(i) The supervisor reports the message $\hat{\theta}^S(\theta_1^A) = \hat{\theta}^S(\theta_2^A) = \ell_2^A$, the agent reports $\hat{\theta}^A(\theta_1^A) = \hat{\theta}^A(\theta_2^A) = \hat{\theta}_2^A$ and the supervisor pays $\xi \in (0, \kappa)$ to the agent.

(ii) The supervisor reports the message $\hat{\theta}^S(\theta_2^A) = \ell_2^A$, the θ_2^A agent reports $\hat{\theta}^A(\theta_2^A) = \theta_2^A$ and the supervisor pays $\xi \in (0, \kappa)$ to this type of agent. The θ_1^A agent does not participate in collusion.

Proof. Consider first the collusive agreement described in (i) above: $\hat{\theta}^S(\theta_1^A) = \hat{\theta}^S(\theta_2^A) = \ell_2^A$, $\hat{\theta}^A(\theta_1^A) = \hat{\theta}^A(\theta_2^A) = \hat{\theta}_2^A$ and the supervisor pays $\xi \in (0, \kappa)$ to the agent. The θ_1^A agent's highest payoff ($\xi = \kappa$) if he participates in collusion is

$\tilde{w} - d(x_2 - \theta_1^A) + \kappa = \bar{w} + d(x_2 - \theta_2^A) - d(x_2 - \theta_1^A) < \bar{w}$. Hence, the θ_1^A agent does not participate in such a collusion agreement since he can guarantee himself a payoff of $\bar{w}$ by doing so.

Consider now the collusive agreement described in (ii) above: $\hat{\theta}^S(\theta_2^A) = \ell_2^A$, $\hat{\theta}^A(\theta_2^A) = \theta_2^A$, the supervisor pays $\xi \in (0, \kappa)$ to the θ_2^A agent and the θ_1^A agent does not participate in collusion. The θ_2^A agent's highest payoff ($\xi = \kappa$) if he participates in collusion is $\tilde{w} - d(x_2 - \theta_1^A) + \kappa = \bar{w}$. This implies that the θ_2^A agent does not participate in such a collusion agreement since he can guarantee himself the same payoff of $\bar{w}$ by doing so.

Proof of Proposition 3:
Notice that given contract CBP there only exist gains-from-trade from collusion in two separate instances. When S observes θ_2^S the type θ_2^A of A is willing to pay up to b as in Condition (12) for S to report $\hat{\theta}_1^S$ rather than $\hat{\theta}_2^S$. When S reports message ℓ_2^A she is willing to pay up to the amount κ for the agent to report $\hat{\theta}_2^A$ rather than $\hat{\theta}_1^A$ or $\mathcal{B}$.[33]

Lemma 5 and Lemma 6 show that neither type of collusion may occur in any equilibrium of our model under contract CBP. In the absence of collusion, Lemma 3 and Lemma 4 show that under contract CBP no blackmail on the part of the agent or of the supervisor will occur in equilibrium. It then follows that, under contract CBP, all the PBE of the continuation game between the supervisor and the agent coincides with the PBE of the corresponding continuation game under the *honest* contract.

References

Aghion, P. and B. Caillaud. 1988. "On the Role of Intermediaries in Organizations". In: *Three Essays in Contract Theory: On the Role of Outside Parties in Contractual Relationships*. by B. Caillaud, Ph.D. thesis, MIT.

Aubert, C., P. Rey, and W. Kovacic. 2006. "The Impact of Leniency and Whistleblowing Programs on Cartels". *International Journal of Industrial Organization* 21: 1241–66.

Austen-Smith, D. and T. J. Feddersen. 2009. "Public Disclosure, Private Revelation or Silence: Whistleblowing Incentives and Managerial Policy". *mimeo*, MEDS Department, Kellogg School of Management.

Baliga, S. and T. Sjöström. 1998. "Delegation and Collusion". *Journal of Economic Theory* 83: 196–232.

Beim, D., A. V. Hirsch, and J. P. Kastellec. 2014. "Whistleblowing and Compliance in the judicial Hierarchy". *The American Journal of Political Science* 58(4): 904–18.

[33] In the latter case S is willing to pay more: $\kappa + \varepsilon$.

Celik, G. 2009. "Mechanism Design with Collusive Supervision". *Journal of Economic Theory* 144: 69–95.

Che, Y.-K. and J. Kim. 2006. "Robust Collusion-Proof Implementation". *Econometrica* 74(4): 1063–107.

De La O, A. and F. Martel Garcia. 2015. "Can Intrastate Accountability Reduce Local Capture? Results from a Field Experiment in Mexico". *mimeo*, Yale University.

Demsky, J. S. and D. E. M. Sappington. 1989. "Hierarchical Structure and Responsibility Accounting". *Journal of Accounting Research* 27: 40–58.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort. 1999. "The Endogenous Transaction Costs of Delegated Auditing". *European Economic Review* 43: 1039–48.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort. 2000. "A Theory of Supervision with Endogenous Transaction Costs". *Annals of Economics and Finance* 1: 231–63.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort. 2001. "On Some Agency Costs of Intermediated Contracting". *Economics Letters* 71: 75–82.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort. 2003. "Collusion, Delegation and Supervision with Soft Information". *Review of Economic Studies* 70: 253–79.

Felli, L. 1990. "Three Essays in Economic Theory: Collusion, Delegation and Search". *PhD thesis*. MIT.

Felli, L. and R. Hortala-Vallve. 2011. "Avoiding Collusion Through Discretion". *mimeo*, London School of Economics.

Fudenberg, D. and J. Tirole. 1991. "Perfect Bayesian Equilibrium and Sequential Equilibrium". *Journal of Economic Theory* 53: 236–60.

Gambetta, D. and P. Reuter. 1995. "Conspiracy among the Many: The Mafia in Legitimate Industries". In: *The Economics of Organised Crime,* Gianluca Fiorentini and Sam Peltzman (eds.) Cambridge: Cambridge University Press.

Hindriks, J., M. Keen, and A. Muthoo. 1999. "Corruption, Extortion and Evasion". *Journal of Public Economics* 74: 395–430.

Kessler, A. S. 2000. "On Monitoring and Collusion in Hierarchies". *Journal of Economc Theory* 91(2): 280–91.

Khalil, F., J. Lawarrée, and S. Yun. 2010. "Bribery versus Extortion: Allowing the Lesser of Two Evils". *RAND Journal of Economics* 41(1): 179–98.

Kofman, F. and J. Lawarrée. 1993. "Collusion in Hierarchical Agency". *Econometrica* 61: 629–56.

Kofman, F. and J. Lawarrée. 1996. "A Prisoner's Dilemma Model of Collusion Deterrence". *Journal of Public Economics* 59: 117–36.

Laffont, J.-J. and D. Martimort. 1997. "Collusion under Asymmetric Information". *Econometrica* 65: 865–912.

Laffont, J.-J. and D. Martimort. 1999. "Collusion-Proof Samuelson Conditions for Public Goods". *Journal of Public Economic Theory* 1(4): 399–438.

Laffont, J.-J. and D. Martimort. 2000. "Mechanism Design with Collusion and Correlation". *Econometrica* 68: 309–42.

Laffont, J. and D. Martimort. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton: Princeton University Press.

Laffont, J.-J. and J. Tirole. 1988. "The Dynamics of Incentive Contracts". *Econometrica* 56: 1153–75.

Leppamaki, M. 1997. "An Economic Theory of Collusion, Blackmail and Whistle-Blowing in Organisations". *PhD thesis*. LSE.

Ma, C.-t., J. Moore, and S. Turnbull. 1988. "Stopping Agents from 'Cheating'". *Journal of Economic Theory* 46: 355–72.

McCubbins, M. and T. Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols Versus Fire Alarms". *American Journal of Political Science* 28: 165–79.

Miceli, M. P., J. P. Near, and T. M. Dworkin. 2008. *Whistle-Blowing in Organizations*. New York: Routledge.

Mookherjee, D. and S. Reichelstein. 1990. "Implementation via Augmented Revelation Mechanisms". *Review of Economic Studies* 57: 453–75.

Motta, M. and M. Polo. 2003. "Leniency Programs and Cartel Prosecution". *International Journal of Industrial Organization* 21: 347–79.

Polinsky, A. M. and S. Shavell. 2001. "Corruption and Optimal Law Enforcement". *Journal of Public Economics* 81: 1–24.

Quesada, L. 2005. "Collusion as an Informed Principal Problem". *mimeo*, Universidad Torcuato Di Tella.

Spagnolo, G. 2004. "Divide et Impera: Optimal Leniency Programs". CEPR Discussion Paper No. 4840.

Spagnolo, G. 2008. "Leniency and Whistleblowers in Antitrust". In: *Handbook on Antitrust Economics,* P. Buccirossi (ed.) Cambridge: MIT Press. Chap. 12.

Ting, M. M. 2008. "Whistleblowing". *The American Political Science Review* 102(2): 249–67.

Tirole, J. 1986. "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations". *Journal of Law, Economics and Organizations* 2: 181–214.