Making Collusion Hard: Asymmetric Information as a Counter-Corruption Measure

Juan Ortner Sylvain Chassang

Boston University Princeton University[*]

March 11, 2014

Preliminary – Do not quote, Do not circulate

Abstract

We study the problem of a principal who relies on the reports of a monitor to provide incentives to an agent. The contracting difficulty that the principal faces is that the monitor and the agent can collude. We show that the principal can improve incentive provision by randomizing over the contract he offers to the monitor, and by informing only the monitor about the realization of this randomization. This randomization adds incomplete information between the monitor and the agent, making collusion between them harder. We characterize the optimal randomized contracts by the principal and quantify the potential gains from using random incentive schemes versus deterministic ones. In practice, a principal can implement this endogenous adverse selection by hiring different monitors and offering them different compensation schemes, and by randomly assigning the monitors to the agents that need to be monitored.

Keywords: corruption, monitoring, collusion, random contracts.

[*] Ortner: jortner@bu.edu, Chassang: chassang@princeton.edu.

1 Introduction

This paper studies how random incentive schemes can temper the negative effects of collusion in organizations. We consider a principal who relies on the reports of a monitor to provide incentives to an agent. We show that the principal can improve incentive provision by randomizing over the incentive contract that he offers the monitor, and by informing only the monitor about the realization of her contract. These randomizations add asymmetric information between the monitor and the agent, making it harder for them to reach a collusive agreement. We derive the optimal randomized contracts by the principal and quantify the potential gains from using random incentive schemes relative to deterministic ones. We also explore practical ways in which a principal can implement this endogenous adverse selection.

The agent in our model decides whether to take a "good" action or a "bad" action, which is unobservable to the principal. Taking the bad action provides a private benefit to the agent, but it inflicts a cost on the principal. The principal hires a monitor, who perfectly observes the agent's action and who sends a report to the principal. With a small probability, the principal receives additional information that he can use to assess the validity of the monitor's report. This can represent other sources of information that the principal might have access to, or can represent information that the principal might obtain by double-checking the agent's report. The principal decides whether or not to punish the agent based on the monitor's report and the additional information she might receive. The principal offers a wage contract to the monitor to incentivize truthful reporting.

The contracting difficulty that the principal faces is that the monitor and the agent may collude: after the agent chooses her action, the monitor and the agent can write a side-contract stipulating the message that the monitor sends to the principal and a transfer from the agent to the monitor.[1] Our model allows agents to have private information about the benefit they derive from taking the bad action, and allows also for different levels of

[1] Therefore, collusion in our model occurs ex-post, after the agent takes her action. In an extension we consider a setting in which the monitor and the agent can collude ex-ante, before the agent takes the action.

bargaining power between the monitor and the agent at the side-contracting stage.

For a concrete example, suppose the agent is a firm that can choose between using a low cost technology that produces high levels of pollution or a more expensive clean technology. The principal is a regulatory agency whose goal is to maintain low levels of contamination. The principal hires a monitor to audit the firm and check its pollution level. The risk that the principal faces is that a corrupt firm may bribe the monitor in exchange of a favorable report. Other examples may include hygiene inspections to restaurants, financial auditing, or monitoring of employees within a firm.

Our analysis highlights how random incentives can help deter corruption by making collusion harder. Suppose the principal randomizes over the incentive scheme that he offers to the monitor, and informs only the monitor the outcome of this randomization. Since monitors with more high-powered incentives face a higher opportunity cost from colluding, this randomization by the principal makes a misbehaving agent uncertain about the cost of bribing the monitor. We show how the principal can leverage this endogenous asymmetric information to reduce the cost (in terms of wage payments to the monitor) of providing incentives to the agent. We derive the optimal randomization by the principal, and illustrate through numerical examples that the potential gains from using random incentive schemes can be substantial.

One potential concern with the mechanism we propose in this paper is that the monitor may be able to signal her incentive contract to the agent and undo the endogenous asymmetric information that the principal generates. We show that the endogenous adverse selection is robust to signaling: the randomization by the principal maintains its incentive properties even if, at the side-contracting stage, the agent can offer the monitor any incentive compatible and individual rational mechanism to extract her (endogenous) private information.

Under the optimal randomization, the principal offers the monitor a wage contract drawn from a distribution whose support has a continuum of wages. We show that the principal can achieve a substantial fraction of the gains from random incentive schemes by using a simpler distribution with only two wage contracts in its support. In practice, if there are many agents that need to be monitored, the principal can implement this two-wage randomization by hiring two types of monitors: a fraction of "elite" monitors who have high-powered incentives, and a fraction of "normal" monitors with low-powered incentives. The principal can then generate asymmetric information at the side-contracting stage by randomly assigning different monitors to different agents.

Finally, the optimal randomization by the principal depends on the agent's bargaining power at the side-contracting stage vis-a-vis the monitor. We show how the principal can design random incentive schemes that have good incentive properties regardless of the relative bargaining power between the agent and monitor. In this way, a principal who is uncertain about the details of how bargaining will take place at the side-contracting stage can still benefit from using randomized incentive schemes.

Our paper is relates primarily to the literature on collusion in mechanism design (i.e., Tirole (1986), Laffont and Martimort (1997, 2000), Baliga and Sjöström (1998), Felli and Villa-Boas (2000), Faure-Grimaud et al. (2003), Mookherjee and Tsumagari (2004), Che and Kim (2006), Celik (2009)). Papers in this literature typically assume that agents have common knowledge about the contracts that the principal offers to each of them. Our point of departure from this literature is the observation that, when there is a threat of collusion, a principal can benefit from maintaining different agents asymmetrically informed about the incentive schemes in place. In our model, the principal achieves this by randomizing over the contracts he offers to the monitor, and by informing only the monitor about the realization of this randomization.

Baliga and Sjöström (1998) also analyze the role of random contracts in agency models with collusion. They consider a setting in which agents face limited liability constraints at the side-contracting stage, so any side-payment that an agent pays must come from the wage she gets from the principal. Baliga and Sjöström (1998) show that a principal can deter collusion by randomizing ex-post over the agents' wages: if the principal pays wages of zero with positive probability, an agent can always claim she has received such a wage and is therefore unable to make a side-payment. The role of randomizations in Baliga and Sjöström (1998) is different than in our setting. The randomizations in their model are ex-post, after the agents collude, so they don't add asymmetric information at the side-contracting stage. Instead, these randomizations interact with the limited liability constraints so as to make it impossible for agents to commit to paying any side-transfers they had agreed to ex-ante.

Our paper relates to Rahman (2012) and Chassang and Padró i Miquel (2013), who also highlight the role of random incentives in a principal-monitor-agent setting. In Rahman (2012), the randomizations allow the principal to jointly provide incentives to the agent and to the monitor who is in charge of supervising the agent. In Chassang and Padró i Miquel (2013), the randomizations allow the principal to garble the information that the monitor provides so as to reduce the agent's incentives to retaliate against the monitor. In contrast to these papers, monitoring is costless in our setting and there is no threat of retaliation. Instead, the randomizations by the principal introduce asymmetric information between the agent and the monitor, and make collusion between them harder.[2]

Finally, other papers have also considered randomized mechanisms that endogenously generate asymmetric information. Calzolari and Pavan (2006a) consider the problem of a monopolist who expects her buyer to resell the good she is purchasing. They show that the optimal selling mechanism is stochastic, with the monopolist selling to different types of buyers with different probabilities. Such a stochastic mechanism allows the monopolist to manipulate the beliefs of the third party regarding the buyers' valuation, and induces this third party to make higher offers at the resale stage.[3] Fudenberg and Tirole (1990) and Ma (1991) study agency models in which the principal can renegotiate the terms of the agent's contract after the agent provides effort. They show that the optimal contract may induce

[2] Strausz (2006), Lazear (2006), Eeckhout et al. (2010), Jehiel (2012), Ederer et al. (2013) and Rahman and Obara (2010) have also emphasized the usefulness of random incentives.

[3] A similar stochastic mechanism is present in Calzolari and Pavan (2006b).

the agent to randomize over the effort she exerts. This randomization generates asymmetric information between the principal and the agent and relaxes renegotiation-proof constraints.

The remainder paper is organized as follows. Section 2 introduces the framework that we study. Section 3 illustrates the main points of our analysis using a simple example. Section 4 studies our general framework and presents our main results. Section 5 extends our analysis to settings in which the agent has two-dimensional private information. Section 6 concludes. Appendix A presents several extensions. Omitted proofs are contained in Appendix B.

2 Framework

The paper explores different cases, depending on the degree of asymmetric information between players, or their ability to commit. They all share the following common structure.

Players and actions. We consider a game with three players: a principal, an agent and a monitor. The agent takes an action $e \in \{0, 1\}$. The agent obtains a payoff $e \times \alpha$ if she takes action e, where $\alpha > 0$ is the private benefit that she gets from taking action $e = 1$. Our general model in Section 4 allows α to be private information of the monitor, drawn from a distribution with support $[\underline{\alpha}, \overline{\alpha}]$. The principal obtains a payoff of $-e \times \beta$ if the agent takes action $e \in \{0, 1\}$, where $\beta > 0$ measures the cost that the principal incurs if the agent takes action $e = 1$. The agent's action is unobservable to the principal.

As an example, suppose the agent is a firm that can choose between using a low cost technology that produces high levels of pollution or a more expensive clean technology, and the principal is a regulatory agency whose goal is to maintain low levels of contamination. In this setting, the parameter α measures the savings from choosing the low cost technology, and β is the cost that the principal incurs when a firm pollutes.

The principal hires the monitor to audit the agent. The monitor observes the agent's action perfectly and at no cost. After observing the agent's action, the monitor sends a message $m \in \{0,1\}$ to the principal. The principal can detect false messages with probability $q \in (0,1)$: if the monitor sends a false message $m \neq e$, the principal observes signal $s = e$ with probability q and observes signal $s = m$ with probability $1 - q$; if the monitor sends a truthful message $m = e$, the principal observes signal $s = m$ with probability 1. For instance, in the case of a firm that may pollute the environment, the regulatory agency might obtain information regarding the firm's actions from other sources such as complaints by employees/locals, an audit by a different governmental agency or through the press.

The principal relies on the monitor's message and his own information to provide incentives to the agent. To induce her to reveal the information she acquired, the principal offers the monitor a wage contract $w = w(m,s)$. The wage is made contingent on the monitor's message and on the principal's information: $w(m,s)$ is the wage in case the monitor sends message $m \in \{0,1\}$ and the principal observes signal $s \in \{0,1\}$. The monitor is protected by limited liability, so $w(m,s) \geq 0$ for all m, s. Throughout the paper we focus on *efficiency wage* contracts: under an efficiency wage contract, the principal pays the monitor a wage $w \geq 0$ when the signal he observes matches the message (i.e., when $s = m$), and pays the monitor a wage of 0 when the signal reveals that the monitor had lied (i.e., when $s \neq m$). An efficiency wage contract is therefore fully characterized by a wage $w \geq 0$. We focus on efficiency wage contracts because they allow us to illustrate the value of endogenous asymmetric information in a cleaner way. In Appendix A.1 we show how our results extend when we allow the principal to offer general wage contracts $w = w(m,s)$.

The principal can punish the agent based on the monitor's message and on his own signal. The agent incurs a cost $t > \overline{\alpha}$ when the principal punishes her. The assumption that $t > \overline{\alpha}$ guarantees that the principal can incentivize all types of agents to take the right action if the monitor always sends a truthful message. The principal punishes the agent if the monitor sends a message indicating that the agent took action $e = 1$ and the principal gets a signal indicating that the monitor's message is truthful (i.e., if $m = s = 1$) or if the monitor sends a message indicating that the agent took action $e = 0$ but the principal gets a signal indicating that the message is false (i.e., if $m = 0$ and $s = 1$). Otherwise, if $m = s = 0$ or if $m = 1 \neq s = 0$, the principal doesn't punish the agent.

Finally, we assume that $\overline{\alpha} > qt$. This assumption puts a bound on the cost that the agent incurs when punished and implies that the principal cannot provide incentives to all types of agent using only his own signal: if this condition did not hold, the principal would be able to provide incentives to the agent to take action $e = 0$ by punishing her whenever he receives a signal $s = 1$, regardless of the monitor's report. Therefore, in this case the principal would not need to incentivize the monitor.[4]

Timing. The timing of the game is as follows:

1. The principal commits to paying the monitor an efficiency wage w drawn from some distribution with support in $[0, \infty)$. The principal's strategy (i.e., the distribution from which the principal draws the monitor's wage) is common knowledge. After drawing the wage, the principal informs the monitor (and not the agent) about its realization.

2. The agent chooses an action $e \in \{0, 1\}$, which is observed only by the monitor.

3. The agent and the monitor write a side-contract. At the side-contracting stage, one player (either the agent or the monitor) is randomly chosen to make a take-it-or-leave-it offer; $\lambda \in [0, 1]$ denotes the probability with which the monitor is selected to make the offer. The offer stipulates the message that the monitor will send to the principal and a transfer from the agent to the monitor. If the other player accepts the offer, the monitor commits to send the message that the side-contract stipulates. Otherwise, the monitor sends the message that maximizes her expected compensation.[5]

[4] The assumption that punishments are bounded may come from a limited liability constraint for the agent. Also, in a richer model in which the monitor's observations are noisy and in which an agent who takes the right action gets punished with positive probability, bounded punishments would arise endogenously if the agent has a participation constraint.

[5] In Appendix A.2 we show how our results extend when, instead of making take-it-or-leave-it offers, at the side-contracting stage the agent and the monitor can use an efficient incentive compatible and individually rational bilateral bargaining mechanism.

4. The monitor sends message $m \in \{0, 1\}$ to the principal. The principal receives the monitor's message and his own signal. He then pays the monitor and punishes the agent based on the monitor's message and his own signal.

Payoffs. If an agent with type α takes action $e = 1$, pays a transfer T to the monitor, and the monitor sends message $m \in \{0, 1\}$ to the principal, her payoffs are

$$U_A = \alpha - T - t \times [\mathbf{1}_{\{m=1\}} + q \times \mathbf{1}_{\{m=0\}}],$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. The agent earns a benefit α by taking action $e = 1$ and pays T to the monitor. By taking action $e = 1$, the agent gets punished with probability 1 if the monitor sends a truthful message $m = 1$, and she gets punished with probability q if the monitor lies and sends message $m = 0$. If the agent takes action $e = 0$, pays a transfer T to the monitor, and the monitor sends message $m \in \{0, 1\}$, her payoffs are

$$U_A = -T - t \times (1 - q) \times \mathbf{1}_{\{m=1\}}.$$

In this case the agent doesn't get punished if the monitor sends a truthful message $m = 0$, and gets punished with probability $1 - q$ if the monitor lies and sends message $m = 1$.

Suppose the agent takes action $e \in \{0, 1\}$ and pays a transfer T to the monitor. Then, the monitor's payoff from sending message $m \in \{0, 1\}$ when her wage is w is

$$U_M = T + w \times (1 - q \times \mathbf{1}_{\{m \neq e\}}).$$

The monitor receives a transfer T and receives a wage w if he sends message $m = e$ or if he sends message $m \neq e$ and the principal doesn't detect that the message is false.

Finally, the principal's payoff if she pays the monitor a wage w, if the agent takes action

$e \in \{0,1\}$ and if the monitor sends message $m \in \{0,1\}$ is

$$U_P = -\beta \times e - w(1 - q \times \mathbf{1}_{\{m \neq e\}}).$$

The principal incurs a loss of $\beta > 0$ when the agent takes action $e = 1$ and pays the monitor a wage w only if he doesn't detect a lie.

All three players are risk-neutral expected utility maximizers.

3 A motivating example

The premise of our work is that to have a durable effect, counter-corruption schemes must be robust to collusion between agents and monitors. Using a detailed example, this section delineates our main argument and suggests that: 1) collusion can severely undermine the efficiency of deterministic counter-corruption schemes, including schemes using several monitors for cross validation; 2) using reward structures that foster asymmetric information between the agent and the monitor may temper the effects of collusion; 3) this endogenous adverse selection is robust to arbitrary signaling by the monitor.

For the sake of simplicity, it is assumed throughout this section that payoff parameters (α, β) are common knowledge among players, with $\alpha \in (qt, t)$, and that the agent has all the bargaining power at the side-contracting stage, i.e. the agent makes the monitor a take-it-or-leave-it offer.

3.1 Benchmark counter-corruption schemes

To frame ideas, we briefly discuss the effectiveness of prominent deterministic counter-corruption schemes depending on whether the agent and the monitor can engage in collusion.

No collusion. Suppose first that the monitor and the agent cannot collude. Note that in this case the principal can induce the monitor to send a truthful message by offering her an efficiency wage $w = 0$. Since the monitor always reports truthfully, the agent would get a payoff of $\alpha - t < 0$ if she takes action $e = 1$ and would get a payoff of 0 if she takes action $e = 0$. Therefore, in the absence of collusion the principal can provide incentives at no cost.

Collusion with deterministic contracts. Consider next a setting in which the monitor and the agent can collude, and in which the principal pays the monitor a deterministic efficiency wage $w \geq 0$.

Suppose the agent takes action $e = 1$ and makes an offer T to the monitor in exchange of a message $m = 0$. The monitor's payoff from accepting this offer is $T + (1 - q)w$, since with probability q the principal detects that the message is false and pays her zero. If the monitor rejects the offer, she finds it optimal to send the truthful message $m = 1$ to the principal and receives a wage w. Therefore, the monitor accepts an offer T if and only if $T \geq qw$.

The agent's payoff from taking action $e = 1$ and making an offer $T < qw$ is then equal to $\alpha - t < 0$: the monitor rejects such an offer, sends a truthful message $m = 1$ to the principal and the agent gets punished with probability 1. On the other hand, the agent's payoff from making an offer $T \geq qw$ is equal to $\alpha - T - qt \leq \alpha - qw - qt$: in this case the monitor sends a favorable message $m = 0$ to the principal and the agent only gets punished if the principal detects that the message is false.

Finally, note that the agent's payoff from taking action $e = 0$ is equal to 0, since in this case the monitor has an incentive to send a truthful message $m = 0$ to the principal even if the agent offers a transfer $T = 0$. Therefore, the agent takes action $e = 0$ if and only if $0 \geq \alpha - qw - qt$: in order to induce the agent to take action $e = 0$, the principal must pay the monitor an efficiency wage $w = \alpha/q - t > 0$ (where the inequality follows since $\alpha > qt$).

Multiple monitors. An alternative approach that has been proposed in the literature (and is often used in practice) to fight collusion is for the principal to hire multiple monitors. Although this can be a practically useful way of fighting collusion and corruption, hiring multiple monitors is also costly for the principal. As the following example illustrates, the cost of fighting collusion and corruption with multiple monitors need not be lower than the cost of fighting collusion with a single monitor.

Suppose that the principal hires two monitors to check the agent's action. Let w_1 and w_2 be the deterministic efficiency wages of monitors 1 and 2, respectively. The timing of the game is as follows: the agent first takes action $e \in \{0, 1\}$. The two monitors then independently observe the agent's action, and each of them independently sends a message $m \in \{0, 1\}$ to the principal. The principal detects whether either one or both messages are false with probability q.[6] The principal punishes the agent either if he receives one message $m = 1$ and doesn't detect that the message is false, or if he receives two messages $m = 0$ and detects that they were false.

The agent can guarantee herself a payoff of zero by taking action $e = 0$, since in this case both monitors would find it optimal to send messages $m = 0$ even without a transfer. On the other hand, if the agent takes action $e = 1$ she has to bribe both monitors to avoid being punished. Monitor $i = 1, 2$ accepts a bribe T if and only if $T \geq qw_i$, so the cost of bribing the two monitors is $q(w_1 + w_2)$. The agent's payoff from taking action $e = 1$ and bribing the two monitors is $\alpha - q(w_1 + w_2) - qt$, so the agent takes action $e = 0$ if and only if $w_1 + w_2 \geq \alpha/q - t$. Note that the cost of providing incentives to the agent in this setting is the same as the cost of providing incentives in a setting with a single monitor who gets a deterministic efficiency wage.

[6] Since monitors observe the agent's action perfectly, if the principal detects that a monitor's message is false then he knows whether the other monitor's message was true or false.

3.2 Adverse selection as a counter-corruption device

We now illustrate how random contracts can improve incentive provision when there is a threat of collusion. Suppose that the principal hires a single monitor and randomizes over the efficiency wage he offers her: with probability $x \in [0, 1]$ the principal offers the monitor a wage $w = 0$, and with probability $1 - x$ the principal offers the monitor a wage $w = \alpha/q - t$. The principal informs only the monitor about the realization of the wage. This randomization by the principal creates asymmetric information between the monitor and the agent, since now the agent is uncertain about the whether the monitor's wage is $w = 0$ or $w = \alpha/q - t$.

Suppose the agent takes action $e = 1$ and makes an offer T to the monitor. The monitor will accept this offer if only if $T \geq qw$. An offer $T \geq \alpha - qt$ will then be accepted by the monitor regardless of whether her wage is $w = 0$ or $w = \alpha/q - t$. The agent's payoff from making such an offer is $\alpha - qt - T \leq 0$. On the other hand, an offer $T \in [0, \alpha - qt)$ will be accepted only by a monitor with wage $w = 0$. The agent's payoff from making an offer $T \in [0, \alpha - qt)$ is then equal to $\alpha - x(T + qt) - (1 - x)t \leq \alpha - xqt - (1 - x)t$: with probability x the offer is accepted and the agent gets a payoff of $\alpha - qt - T$, and with probability $1 - x$ the offer is rejected and the agent gets a payoff of $\alpha - t$.

The agent's payoff from taking action $e = 0$ is equal to 0, since in this case the monitor has an incentive to send a truthful message $m = 0$ even in the absence of collusion. Therefore, the agent will take action $e = 0$ if and only if $\alpha - xqt - (1 - x)t \leq 0$, or $x \leq \frac{t-\alpha}{(1-q)t}$. The assumption that $\alpha \in (qt, t)$ implies that $\frac{t-\alpha}{(1-q)t} \in (0, 1)$. It follows that the principal can provide incentives to the agent by paying the monitor a wage $w = 0$ with probability $x = \frac{t-\alpha}{(1-q)t}$ and paying her a wage $w = \alpha/q - t$ with probability $1 - x = \frac{\alpha-qt}{(1-q)t}$. The expected wage that the principal pays the monitor under this randomization is $\frac{\alpha-qt}{(1-q)t}(\frac{\alpha}{q} - t)$, which is a fraction of the deterministic wage $w = \alpha/q - t$ derived above.

Intuitively, when the principal randomizes over the monitor's wage, the agent is uncertain about the type of monitor she is facing: she could be facing a low-paid monitor who is easy to corrupt, or a high-paid monitor who has a higher cost of accepting a bribe. Given this uncertainty, the principal no longer needs to pay the monitor a high wage all the time: the agent will still have an incentive to take the right action if the probability of encountering a high-paid monitor is large enough.

In practice, if there are many agents that need to be monitored, the principal can implement this two-wage randomization by hiring two types of monitors: a fraction $1 - x$ of “elite” monitors, who have a high efficiency wage equal to $w = \alpha/q - t$, and a fraction x of “normal” monitors, who have a low efficiency wage of $w = 0$. The principal can then generate asymmetric information at the side-contracting stage by randomly assigning different monitors to different agents.

3.3 Robustness to signaling

A potential concern with the randomization in Section 3.2 is that the monitor may be able to signal her wage to the agent and undo the endogenous asymmetric information that the principal creates. We now show that the endogenous adverse selection is robust to signaling: the agent would still have an incentive to take action $e = 0$ even if, instead of making a take-it-or-leave-it offer at the side-contracting stage, she could design any incentive compatible and individually rational mechanism to extract the monitor’s private information.

Suppose that the principal uses the randomization in Section 3.2: he offers the monitor a wage $w = 0$ with probability $x = \frac{t-\alpha}{(1-q)t}$ and a wage $w = \overline{w} = \alpha/q - t$ with probability $1-x = \frac{\alpha-qt}{(1-q)t}$. If an agent takes action $e = 1$, then at the side-contracting stage she can offer the monitor any incentive compatible and individually rational bargaining mechanism. By the revelation principle we can focus on mechanisms in which the monitor reports her wage truthfully. A bargaining mechanism is characterized by four quantities $(P_0, T_0, P_{\overline{w}}, T_{\overline{w}})$: P_w is the probability that the monitor and the agent reach an agreement when the monitor’s wage is w, and T_w is the expected transfer from the agent to a monitor with wage w.

Under mechanism $(P_0, T_0, P_{\overline{w}}, T_{\overline{w}})$, the agent's payoff from taking action $e = 1$ when the monitor's wage is w is $\alpha - T_w + P_w(-qt) + (1 - P_w)(-t) = \alpha - t + P_w(1 - q)t - T_w$. The agent earns α from taking action $e = 1$ and pays the monitor an expected transfer T_w. With probability P_w the agent and the monitor reach an agreement, the monitor sends message $m = 0$ to the principal, and the principal punishes the agent only if he detects that the message is false. With probability $1 - P_w$ there is no agreement, the monitor sends message $m = 1$ and the agent gets punished with probability 1. The agent's expected payoff from taking action $e = 1$ and offering the monitor a mechanism $(P_0, T_0, P_{\overline{w}}, T_{\overline{w}})$ is then given by $\alpha - t + x(P_0(1 - q)t - T_0) + (1 - x)(P_{\overline{w}}(1 - q)t - T_{\overline{w}})$.

Under mechanism $(P_0, T_0, P_{\overline{w}}, T_{\overline{w}})$, a monitor with wage w who reports that her wage is w' obtains a payoff $U_M(w, w') = T_{w'} + P_{w'}(1 - q)w + (1 - P_{w'})w = T_{w'} + w(1 - qP_{w'})$. The monitor gets a transfer $T_{w'}$. If there is no agreement, the monitor sends a truthful message to the principal and gets her wage w. Otherwise, if there is agreement, the monitor sends a message $m = 0$ and gets an expected wage of $(1 - q)w$. By incentive compatibility, it must be that $U_M(w, w) \geq U_M(w, w')$ for $w, w' \in \{0, \overline{w}\}$. By individual rationality, it must be that $U_M(w, w) \geq w$: the monitor's payoff cannot be smaller than her payoff from not participating in the mechanism and sending a truthful message.

The agent's problem is to design a mechanism that maximizes her payoff subject to these constraints. By standard arguments, the constraints that bind in this setting are: (i) the IR constraint of a monitor with wage $\overline{w}$, and (ii) the IC constraint of a monitor with zero wage. Constraint (i) implies that $U_M(\overline{w}, \overline{w}) = \overline{w}$, which is satisfied when $T_{\overline{w}} = P_{\overline{w}}\overline{w}q$, while constraint (ii) implies that $T_0 = T_{\overline{w}}$. Therefore, the agent's payoff is from taking action $e = 1$ and offering an optimal mechanism $(P_0, T_0, P_{\overline{w}}, T_{\overline{w}})$ is $\alpha - t + x(P_0(1 - q)t - P_{\overline{w}}\overline{w}q) + (1 - x)P_{\overline{w}}((1 - q)t - q\overline{w})$. Note that it is always optimal to set $P_0 = 1$ and $P_{\overline{w}}$ equal to either 0 or 1 (since this expression is linear in $P_{\overline{w}}$). If $P_{\overline{w}} = 0$ the agent's payoff is $\alpha - t + x(1 - q)t = 0$, while if $P_{\overline{w}} = 1$ the agent's payoff is $\alpha - t + (1 - q)t - q\overline{w} = 0$. The agent gets a payoff of 0 in either case, which is what she gets by taking action $e = 0$.

4 The General Case

4.1 Framework

The model in this section generalizes the example of Section 3 in three ways. First, we allow agents to have private information about the benefit that they derive from taking action $e = 1$. We assume that the agent's private benefit α is drawn from a distribution with cdf G and with support $[\underline{\alpha}, \overline{\alpha}]$. Recall from Section 2 that the support $[\underline{\alpha}, \overline{\alpha}]$ of possible agent types is such that $\overline{\alpha} \in (qt, t)$ (see Section 2 for a justification of this assumption). Second, we allow situations in which the monitor also has bargaining power: at the side-contracting stage the monitor makes a take-it-or-leave offer to the agent with probability $\lambda \in [0, 1]$ and the agent makes a take-it-or-leave offer to the monitor with probability $1 - \lambda$. Finally, instead of focusing on a two-wage distribution as in Section 3.2, we place no restrictions on the randomization over the monitor's wage that the principal uses. The timing of the game and the players' payoffs are the same as in Section 2.

4.2 Analysis

Incentives for a given wage schedule. Let $\mathcal{F}$ be the set of all cumulative density functions (cdfs) with support on $[0, \infty)$. Suppose that the principal randomizes over the wage he offers to the monitor according to a distribution with cdf $F \in \mathcal{F}$, and consider an agent with type $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ who takes action $e = 1$. Suppose first that the agent is selected to make an offer to the monitor at the side-contracting stage. The monitor's payoff from accepting an offer T is $T + (1 - q)w$, while her payoff from rejecting the offer and sending a truthful message is w. Therefore, a monitor accepts an offer T if and only if $T \geq qw$. From the agent's perspective, the monitor accepts an offer T with probability $F(T/q)$.

Let T be the offer that the agent makes when selected to make an offer after taking action $e = 1$. When the agent takes action $e = 1$ and is selected to make an offer, her payoff is

$\alpha + F(T/q)(-T - qt) + (1 - F(T/q))(-t) = \alpha - t + F(T/q)((1-q)t - T)$. The agent earns α from taking action $e = 1$. If the monitor accepts the offer, she sends message $m = 0$ and the agent gets punished only if the principal detects that the message was false. If the monitor rejects the offer, she sends message $m = 1$ and the agent gets punished with probability 1.

Suppose next that the monitor is selected to make an offer at the side-contracting stage after the agent took action $e = 1$. The agent's payoff from accepting an offer T is $\alpha - T - qt$, since in this case the monitor sends message $m = 0$ and the agent will only get punished if the principal detects that the monitor's message was false. The agent's payoff from rejecting the monitor's offer is $\alpha - t$, since the monitor sends a message $m = 1$ if her offer is rejected. Therefore, the agent will accept the monitor's offer T if and only if $T \leq (1 - q)t$.

The monitor's payoff from making an offer $T = (1-q)t$ that the agent accepts is $(1-q)t + (1 - q)w$, since with probability q the principal detects that the monitor's message is false and pays her a wage of zero. The monitor's payoff from making an offer $T > (1-q)t$ that the agent rejects is w. Hence, the monitor makes an offer $T = (1-q)t$ if $qw \leq (1-q)t$ and makes an offer $T > (1 - q)t$ otherwise. Note that, regardless of whether the monitor makes an offer $T = (1 - q)t$ or an offer $T > (1 - q)t$, the agent's payoff from taking action $e = 1$ when the monitor makes the take-it-or-leave-it-offer is $\alpha - t$: if the monitor makes offer $T = (1 - q)t$ the agent's payoff is $\alpha - qt - T = \alpha - t$, while if the monitor makes offer $T > (1 - q)t$ the agent gets $\alpha - t$ since she gets punished with probability 1. The agent's expected payoff from taking action $e = 1$ and making offer T is equal to $\alpha - t + (1 - \lambda)F(T/q)((1 - q)t - T)$ (where λ is the probability with which the monitor makes the offer).

The payoff that an agent gets by taking action $e = 0$ is 0, since in this case the monitor has an incentive to send a truthful message even if there is no collusion. An agent with type α has an incentive to take action $e = 0$ if and only if,

$$\forall T \geq 0, \;\; \alpha - t + (1 - \lambda)F(T/q)((1 - q)t - T) \leq 0. \tag{1}$$

From equation (1) it follows that if an agent with type α finds it optimal to take action $e = 0$, then all agents with type $\alpha' < \alpha$ will also find it optimal to take action $e = 0$. Similarly, if an agent with type α finds it optimal to take action $e = 1$, then all agents with type $\alpha' > \alpha$ will also find it optimal to take action $e = 1$. By choosing the distribution $F \in \mathcal{F}$ of wages, the principal effectively chooses the cutoff type $\alpha \in [\overline{\alpha}, \underline{\alpha}]$ such that agents with type below α take action $e = 0$ and agents with type above α take action $e = 1$.

Finally, note that the principal can induce all agents with type $\alpha \leq t[1-(1-\lambda)(1-q)]$ to take action $e = 0$ by paying the monitor monitor a wage $w = 0$ with probability 1. In this case, an agent who took action $e = 1$ would offer a transfer $T = 0$ to the monitor when she is selected to make offers and her expected payoff would be $\alpha - t + (1-\lambda)(1-q)t$. Hence, all agents with type $\alpha \leq t[1-(1-\lambda)(1-q)]$ would have an incentive to take action $e = 0$.

Deterministic wages. Consider a principal who wants all agents with type below a cutoff $\alpha > t[1-(1-\lambda)(1-q)]$ to take action $e = 0$. Suppose that the principal pays the monitor a wage w with probability 1. Using (1), the payoff that an agent with type α gets from taking action $e = 1$ and making an offer $T = qw$ when she is selected to make offers is $\alpha - t + (1-\lambda)((1-q)t - qw)$. This agent will have an incentive to take action $e = 0$ only if $\alpha - t + (1-\lambda)((1-q)t - qw) \leq 0$. The lowest efficiency wage that induces agents with types below α to take action $e = 0$ is $\overline{w}_\alpha := \frac{\alpha - t + (1-\lambda)(1-q)t}{q(1-\lambda)}$.

Optimal randomization. Consider next a principal who randomizes over the wage he pays to the monitor. Suppose that the principal wants to provide incentives to agents with type below α to take action $e = 0$, with $\alpha > t - (1-\lambda)(1-q)t$. The principal's problem is to find the cdf in $\mathcal{F}$ that minimizes wage payments subject to the constraint that an agent with type α has an incentive to take action $e = 0$. That is, the principal's problem is

$$\inf_{F \in \mathcal{F}} \int w dF(w) \text{ s.t. } (1) \qquad (2)$$

Proposition 1. *Suppose that the principal wants to provide incentives to all agents with type below α to take action $e = 0$. Then, it is optimal for the principal to pay the monitor a wage drawn from a distribution with cdf $F_\alpha \in \mathcal{F}$ such that, for all $w \geq 0$,*

$$F_\alpha(w) = \min\left\{\frac{t-\alpha}{(1-\lambda)((1-q)t-qw)}, 1\right\}. \tag{3}$$

Proof. We first show that, if F is solves (2), it must be that the constraint (1) holds with equality for all offers T such that $F(T/q) < 1$. To see this, suppose by contradiction that $(1-\lambda)F(T/q)((1-q)t-T) < t-\alpha$ for some $T \geq 0$ such that $F(T/q) < 1$. Let $\tilde{F} \in \mathcal{F}$ be such that, for all T with $\tilde{F}(T/q) < 1$, $(1-\lambda)\tilde{F}(T/q)((1-q)t-T) = t-\alpha$. Note that $\tilde{F}(s) \geq F(s)$ for all s, with strict inequality for some s'; i.e., F first order stochastically dominates $\tilde{F}$. It follows that the expected wage payments under $\tilde{F}$ are strictly lower than under F. Note further that the agent would take action $e = 0$ under distribution $\tilde{F}$, since $(1-\lambda)\tilde{F}(T/q)((1-q)t-T) \leq t-\alpha$ for all $T \geq 0$. This contradicts the assumption that F solves (2). Hence, (1) holds with equality for all $T \geq 0$ such that $F(T/q) < 1$.

Let F be the solution to (1) and let $\overline{T} = \inf\{T : F(T/q) = 1\}$. The arguments in the previous paragraph imply that, for all $T \in [0, \overline{T}]$,

$$(1-\lambda)F(T/q)((1-q)t-T) = t-\alpha \Rightarrow F(T/q) = \frac{t-\alpha}{(1-\lambda)((1-q)t-T)}. \tag{4}$$

Applying the change of variable $w = T/q$ in (4), it follows that the solution to (2) is (3). □

Proposition 1 characterizes the optimal randomization over wages that induces all agents with type $\alpha' \leq \alpha$ to take action $e = 0$. Note that the highest wage in the support of this distribution is $\overline{w}_\alpha = \frac{\alpha-t+(1-\lambda)(1-q)t}{q(1-\lambda)}$, which is the deterministic wage that induces all agents with type below α to take action $e = 1$. Note further that, for all $w \geq 0$, $F_\alpha(w)$ is decreasing in α and is increasing in t and q. Therefore, the cost that the principal has to incur to provide incentives to agents with type lower than α is increasing in α and is decreasing in both t and q.

Finally, note for all $w \geq 0$, $F_\alpha(w)$ is increasing λ. Moreover, for all α there exists $\lambda_\alpha \in (0,1)$ such that $F_\alpha(0) = 1$ for all $\lambda \geq \lambda_\alpha$. This implies that it is cheaper for the principal to provide incentives to the agent when the monitor has more bargaining power at the side-contracting stage; and when the monitor's bargaining power is large enough, the principal can provide incentives to the agent at zero cost (i.e., paying the monitor a wage of zero with probability 1). Intuitively, the monitor extracts more rents at the side-contracting stage as her bargaining power increases. This reduces the payoff that the agent gets by taking action $e = 1$, and makes it cheaper for the principal to provide incentives.

Using the optimal distribution in Proposition 1, the cost in terms of expected wages of providing incentives to agents with type lower than $\alpha > t[1-(1-\lambda)(1-q)]$ is

$$\begin{aligned} \int w dF_\alpha(w) &= \frac{\alpha - t + (1-q)t(1-\lambda)}{q(1-\lambda)} + \frac{t-\alpha}{q(1-\lambda)} \ln\left(\frac{t-\alpha}{(1-\lambda)(1-q)t}\right) \\ &< \frac{\alpha - t + (1-q)t(1-\lambda)}{q(1-\lambda)} = \overline{w}_\alpha, \end{aligned} \tag{5}$$

where the inequality follows since $\alpha > t[1-(1-\lambda)(1-q)]$ implies that $\frac{t-\alpha}{(1-\lambda)(1-q)t} < 1$.

Table 1 compares the cost of providing incentives to agents with type below α for different parameter values under deterministic wages (second column) and under the optimal wage distribution in Proposition 1 (third column). As the table shows, the cost of incentives can potentially be significantly lower with random wage contracts than with deterministic ones.

Optimal cutoffs. Consider the case of a principal who compensates the monitor with a deterministic efficiency wage. Let $\overline{w}_\alpha = \min\{\frac{\alpha-t+(1-q)t(1-\lambda)}{q(1-\lambda)}, 0\}$ be the wage that the principal needs to pay to the monitor to provide incentives to agents with types below α to take action $e = 0$. The principal's payoff from paying the deterministic wage $\overline{w}_\alpha$ is

$$U_P^D(\alpha) := G(\alpha)(-\overline{w}_\alpha) + (1 - G(\alpha))(-\beta - (1-q)\overline{w}_\alpha).$$

Table 1: Cost of Providing Incentives

Parameters	Deterministic wage	Optimal wage distribution	Two-wage distribution
$\alpha = 3, t = 4, q = 0.1, \lambda = 0.3$	21.7	8.5	13.1
$\alpha = 2.5, t = 4, q = 0.1, \lambda = 0.3$	14.6	3.5	5.9
$\alpha = 2, t = 4, q = 0.1, \lambda = 0.3$	7.4	0.8	1.5

Parameters	Deterministic wage	Optimal wage distribution	Two-wage distribution
$\alpha = 2.5, t = 4, q = 0.1, \lambda = 0.2$	17.3	5	8.2
$\alpha = 2.5, t = 4, q = 0.1, \lambda = 0.3$	14.6	3.5	5.9
$\alpha = 2.5, t = 4, q = 0.1, \lambda = 0.4$	11	1.9	3.4

When the agent's type is lower than α, the agent takes action $e = 0$, the monitor sends message $m = 0$, and the principal pays a wage $\overline{w}_\alpha$. When the agent's type is larger than α, the agent takes action $e = 1$ and bribes the monitor, the monitor sends message $m = 0$, and the principal pays a wage $\overline{w}_\alpha$ only if he doesn't detect that the message was false. Therefore, the problem of a principal who compensates the monitor with a deterministic wages is to choose a cutoff $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ to maximize $U_P^D(\alpha)$. Let $\alpha^D \in \arg\max U_P^D(\alpha)$.

Assumption 1. $\alpha^D > t[1 - (1-q)(1-\lambda)]$.

Assumption 1 implies that a principal who compensates the monitor with a deterministic wage would find it optimal to pay a strictly positive wage. Assumption 1 holds when β is large and when the probability that the agent's type is below $t[1 - (1-q)(1-\lambda)]$ is small.

Consider next the case in which the principal randomizes over the wage he pays the monitor. By Proposition 1, the principal will draw the monitor's wage from the distribution with cdf F_α if he wants to incentivize agents with type lower than α to take action $e = 0$. The principal's payoff from using distribution F_α is

$$U_P^R(\alpha) := G(\alpha)\left(-\int w dF_\alpha(w)\right) + (1 - G(\alpha))\left(-\beta - (1-q)\int w dF_\alpha(w)\right).$$

When the agent's type is lower α, the agent takes action $e = 0$, the monitor sends message $m = 0$ and the principal pays the monitor her wage w drawn from the cdf F_α. When the agent's type is larger than α, the agent takes action $e = 1$ and bribes the monitor, the monitor sends message $m = 0$, and the principal pays a wage w only if he doesn't detect that the message was wrong. The principal's problem is to choose $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ that maximizes $U_P^R(\alpha)$.[7] Let α^R be the cutoff that maximizes $U_P^R(\alpha)$.

Proposition 2. *Suppose Assumption 1 holds. Then,* $U_P^R(\alpha_R) > U_P^D(\alpha_D)$.

Proof. Assumption 1 implies that $\overline{w}_{\alpha_D} > 0$. Since $\int w dF_{\alpha_D} < \overline{w}_{\alpha_D}$ (see equation (5)), it follows that $U_P^D(\alpha_D) < U_P^R(\alpha_D) \leq U_P^R(\alpha_R)$. □

Robustness to signaling. As in Section 3, a potential concern with randomized incentive schemes is that the monitor might be able to signal her wage to the agent and undo the asymmetric information that the principal generates. We now show that the endogenous adverse selection is robust to signaling.

To illustrate this point, we assume that an agent who took action $e = 1$ and who is selected to make a proposal at the side-contracting stage can offer the monitor any incentive compatible and individually rational bilateral bargaining mechanism to extract the monitor's private information. We show that an agent with type $\alpha' \leq \alpha$ still has an incentive to take action $e = 0$ if the principal randomizes over the monitor's wage according to F_α.

By the revelation principle we can focus on direct revelation mechanisms in which the agent asks the monitor to report her private information; i.e., her wage. A bilateral bargaining mechanism is characterized by two functions: (i) $P(w)$, the probability with which the monitor and the agent reach an agreement when the monitor's wage is w; and (ii) $T(w)$, the expected transfer from the agent to the monitor when the monitor's wage is w.

[7] When computing $U_P^R(\alpha)$, we assume that an agent who takes action $e = 1$ offers a bribe $T = q\overline{w}_\alpha$ that all monitors accept. From the proof of Proposition 1, it follows that an agent who took action $e = 1$ is indifferent between making any offer $T \in [0, q\overline{w}_\alpha]$. However, the principal is strictly better-off when the an agent who took action $e = 1$ makes offer $T = q\overline{w}_\alpha$, since in this case he won't have to pay the monitor's wage after detecting that the monitor's message was false.

Suppose the principal offers the monitor a wage drawn from F_α and consider an agent with type α who took action $e = 1$. If the monitor is selected to make an offer, then this agent obtains a payoff equal to $\alpha - t$: in this case the monitor finds it optimal to make an offer $T = (1-q)t$ which the agent accepts, and the agent gets $\alpha - qt - (1-q)t = \alpha - t$.[8] If the agent is selected to be proposer and offers the monitor a mechanism (P, T), her payoff is

$$\alpha + \int [-T(w) + P(w)(-qt) + (1 - P(w))(-t)]\, dF_\alpha(w).$$

The agent gets a benefit α from taking action $e = 1$. If the monitor's wage is w, the agent pays an expected transfer $T(w)$ and the monitor and agent reach an agreement with probability $P(w)$. If they reach an agreement the monitor sends message $m = 0$ and the agent gets punished with probability q, and if they don't reach an agreement the monitor sends message $m = 1$ and the agent gets punished with probability 1.

Let $U_A^{\alpha'}(P, T)$ denote the expected payoff that an agent with type α' gets by taking action $e = 1$ and offering mechanism (P, T) to the monitor when selected to be proposer:

$$\begin{aligned} U_A^{\alpha'}(P,T) &= (1-\lambda)\left[\alpha' + \int [-T(w) + P(w)(-qt) + (1-P(w))(-t)]\, dF_\alpha(w)\right] + \lambda(\alpha' - t) \\ &= \alpha' - t + (1-\lambda)\int (P(w)(1-q)t - T(w)) dF_\alpha(w). \qquad (6) \end{aligned}$$

Our next result shows that, for any incentive compatible and individually rational mechanism (P, T) that the agent can offer to the monitor, the expected payoff that an agent with type α gets taking action $e = 1$ is weakly lower than 0, which is the payoff this agent would get from taking action $e = 0$.

Proposition 3. *Suppose that the principal offers the monitor a wage drawn from the distribution F_α. Then, for any incentive incentive compatible and individually rational mechanism*

[8] Note that all wages in the support of F_α are strictly lower than $\frac{(1-q)t}{q}$, so all monitors find it optimal to make offer $T = (1-q)t$ to the agent.

(P, T) and for any $\alpha' \leq \alpha$, $U_A^{\alpha'}(P, T) \leq 0$.

Proof. See Appendix B.1. □

4.3 Practical concerns and the cost of simplicity

Sensitivity to bargaining power. The optimal distribution derived in Proposition 1 depends on the agent's and monitor's bargaining power at the side-contracting stage. In particular, as the agent's bargaining power increases it becomes more costly to provide incentives to the agent. This implies that the worst possible situation for the principal is a setting in which the agent has all the bargaining power (i.e., a setting with $\lambda = 0$).

To highlight the dependency of the optimal distribution on the level of bargaining power, let $F_\alpha(w; \lambda)$ denote the cdf in Proposition 1 when the monitor makes offers with probability λ. Note that for every possible wage distribution $F \in \mathcal{F}$, the agent's payoff $\alpha - t + (1 - \lambda)F(T/q)((1 - q)t - T)$ from taking action $e = 1$ and making offer T is decreasing in λ. Therefore, if the principal pays the monitor a wage drawn from $F_\alpha(w; \lambda)$, the agent will have an incentive to take action $e = 0$ when the level of bargaining power is $\lambda' \in [\lambda, 1]$.

The observations in the two previous paragraphs imply the following result.

Proposition 4. *If the principal pays the monitor a wage drawn from the cdf $F_\alpha(w; 0)$, all agents with type $\alpha' \leq \alpha$ will take action $e = 0$ regardless of the monitor's bargaining power.*

Loss from efficiency wages. A simplification we made so far is to assume that the principal compensates the monitor with an efficiency wage contract: he pays the monitor a wage $w \geq 0$ unless he detects that the monitor's message was false. Efficiency wages not only allow us to illustrate how the principal can benefit from creating endogenous adverse selection in a clean way, but are also easy to implement in practice.

In Appendix A.1 we show how our results extend when we allow the principal to offer general wage contracts. There, we show that the optimal wage contract takes the form of a *bonus contract*: the principal pays the monitor a higher wage when she sends message $m = 1$ than when she sends message $m = 0$. Moreover, we derive the optimal randomization over bonus contracts (i.e., the randomization that minimizes the cost of providing incentives).

How much does the principal lose by using efficiency wages instead of the optimal bonus contract? In Appendix A.1 we show that the cost in terms of expected wages that the principal incurs to provide incentives to agents with a type lower than $\alpha > t[1-(1-\lambda)(1-q)]$ when he uses randomized bonus contracts is equal to $\frac{1-q}{2-q}$ times the cost of providing incentives using randomized efficiency wages. That is, when q is small the expected wage payments are roughly twice as large with efficiency wages than with bonus contracts.

Loss from two-wage distribution. By Proposition 1, under the optimal randomization the principal offers a continuum of wages to the monitor with positive probability. We now quantify the loss that a principal incurs if he instead uses a simpler distribution with two levels of wages in its support.

Consider a principal who wants to provide incentives to agents with type lower than $\alpha > t[1 - (1 - \lambda)(1 - q)]$ to take action $e = 1$. Suppose that the principal offers the monitor a wage drawn the following distribution: with probability $x \in [0, 1]$ the principal offers the monitor a wage $w = 0$, and with probability $1 - x$ her offers a wage $w = \overline{w}_\alpha$.

Under this distribution, an agent who took action $e = 1$ would find it optimal to make either an offer $T = 0$ that only monitors with wage $w = 0$ accept, or an offer $T = q\overline{w}_\alpha$ that all monitors accept. An agent with type α who takes action $e = 1$ and makes an offer $T = q\overline{w}_\alpha$ when selected to make an offer earns an expected payoff of $\alpha - t + (1-\lambda)((1-q)t - q\overline{w}_\alpha) = 0$. On the hand, if this agent instead makes an offer $T = 0$ she earns an expected payoff of $\alpha - t + (1 - \lambda)x(1 - q)t$. An agent with type α has an incentive to take action $e = 0$ (and get a payoff of 0) only if $\alpha - t + (1 - \lambda)x(1 - q)t \leq 0$, or if $x \leq \frac{t-\alpha}{(1-\lambda)(1-q)t}$. Therefore, under this two-wage distribution the expected wage payments that a principal has to make to the monitor to provide incentives to agents with type lower than α are $(1 - \frac{t-\alpha}{(1-\lambda)(1-q)t})\overline{w}_\alpha$.

The following result compares the cost of providing incentives using this two-wage distribution against the cost of providing incentives using the optimal distribution F_α.

Proposition 5. *The cost of providing incentives under the two-wage distribution is at most twice as large as the cost of providing incentives under the distribution in Proposition 1.*

Proof. See Appendix B.1. □

The last column in Table 1 reports the cost of providing incentives (in terms of expected wage payments) under the two-wage distribution for different parameter values. As the table shows, a substantial fraction of the gains from using random wages can be achieved using a more practical two wage distribution.

As mentioned in Section 3.2, if there are many agents that need to be monitored the principal can implement this two-wage randomization by hiring two types of monitors: a fraction $1 - x$ of "elite" monitors, who have a high efficiency wage equal to $\overline{w}_\alpha$, and a fraction x of "normal" monitors, who have a low efficiency wage of $w = 0$. The principal can then generate asymmetric information at the side-contracting stage by randomly assigning different monitors to different agents.

5 Two-dimensional types of agents

The model in the previous sections assumes that all agents incur the same cost t when punished by the principal. Proposition 1 shows that, in this setting, it is optimal to pay the monitor an efficiency wage drawn from a distribution with support $[0, \overline{w}_\alpha]$; that is, it is optimal for the principal to pay a fraction of monitors the lowest possible wage of 0.

This section studies a setting in which agents differ not only on the benefit α that they obtain when taking action $e = 1$, but also in the cost t that they incur when they are punished by the principal. For instance, some agents might have contacts in the government and are therefore less likely to incur any actual punishment. Also, if the punishment t represents a tax that the agent has to pay when punished, agents that are financially constrained would face a higher punishment cost than agents who don't face such constraints. We show that, in this setting, it might be optimal for the principal to offer the monitor a wage drawn from a distribution with support bounded away from zero. To illustrate this point, we consider a simple setting in which agents may have one of two possible levels of punishment cost $t_1 < t_2$. The cost of punishment is also the agent's private information, and is independent of the private benefit α. Let $p \in (0,1)$ be the probability that an agent has cost of punishment t_2.

To see how this additional dimension of uncertainty affects incentives, consider an agent with type (α, t) who takes action $e = 1$, and suppose that the monitor is selected to make an offer at the side-contracting stage. Note that the payoff that the agent gets from accepting an offer T is $\alpha - qt - T$, while her payoff from rejecting the offer is $\alpha - t$. Therefore, an agent accepts an offer T if and only if $T \leq (1-q)t$. When all agents have the same cost t of being punished, the monitor always makes an offer at least as large as $(1-q)t$. This implies that the agent gets no rents from the collusive agreement when the monitor makes the offer.

When the cost that agents incur from being punished can take values t_1 and $t_2 > t_1$, there are three possible optimal offers that a monitor might make at the side-contracting stage: a low offer $T = (1-q)t_1$ that is accepted by all agents, a middle offer $T = (1-q)t_2$ that is accepted only by agents with type (α, t_2), and an offer $T > (1-q)t_2$ that no agent accepts. In Appendix B.2 we show that the optimal offer for a monitor with wage w is

$$T(w) = \begin{cases} (1-q)t_1 & \text{if } w < \frac{1-q}{q(1-p)}(t_1 - pt_2), \\ (1-q)t_2 & \text{if } w \in [\frac{1-q}{q(1-p)}(t_1 - pt_2), \frac{(1-q)}{q}t_2], \\ T > (1-q)t_2 & \text{if } w > \frac{(1-q)}{q}t_2. \end{cases} \tag{7}$$

Intuitively, low paid monitor face a lower cost of colluding, so they find it optimal to make make lower offers that are accepted more frequently.

Let $\hat{w} := \frac{1-q}{q(1-p)}(t_1 - pt_2)$ and assume that $t_1 > pt_2$ (so that $\hat{w} > 0$). Suppose the principal randomizes over the monitor's wage according to $F \in \mathcal{F}$, and let $F(\hat{w}^-) = \lim_{w \uparrow \hat{w}} F(w)$ be the probability that a monitor's wage is strictly lower than $\hat{w}$. In Appendix B.2 we show that the expected payoff that an agent with type (α, t_2) gets when she takes action $e = 1$ is equal to $\alpha - t_2 + (1-\lambda)F(T/q)((1-q)t_2 - T) + \lambda F(\hat{w}^-)(1-q)(t_2 - t_1)$. The last term in this expression represents the payoff that an agent with type (α, t_2) gets if she is facing a monitor with wage lower than $\hat{w}$ and the monitor makes the offer at the side-contracting stage: such a monitor would make a low offer of $(1-q)t_1$, which an agent with type (α, t_2) strictly accepts. Note that this extra payoff that an agent with type (α, t_2) gets is strictly increasing in the fraction $F(\hat{w}^-)$ of monitors that have a wage lower than $\hat{w}$.

Let $\overline{w}_\alpha(t_2) = \frac{\alpha - t_2 + (1-\lambda)(1-q)t_2}{(1-\lambda)q}$. The following result shows that, if the principal wants to provide incentives to agents with type (α, t_2) to take action $e = 1$, it might be optimal to pay the monitor a wage drawn from a distribution with support $[\hat{w}, \overline{w}_\alpha(t_2)]$.

Proposition 6. *Suppose the principal wants to provide incentives to all agents with type (α', t_2) with $\alpha' \leq \alpha$ for some $\alpha > t_2(1 - (1-\lambda)(1-q))$. Then, there exists $\bar{t} \in (pt_2, t_2)$ such that it is optimal for the principal to pay the monitor a wage drawn from the distribution with support $[\hat{w}, \overline{w}_\alpha(t_2)]$ whenever $t_1 \in (pt_2, \bar{t})$.*

Proof. See Appendix B.2. □

6 Discussion

6.1 Summary

A general message from our analysis is that, when there is threat of collusion, a principal can gain from allocating private information differently across agents. This adds asymmetric information between the agents, and makes it harder for them to collude. The way the principal achieves this in our model is by randomizing over the monitor's incentive contract, and by informing only the monitor about the realization of this randomization.

A practical way in which a principal can implement the mechanism we propose in this paper is by offering different incentive schemes to the different monitors that he hires, and by then randomly assigning monitors to audit different agents. This makes the agents uncertain about the type of monitor they are dealing with, and reduces the cost of providing incentives.

6.2 Further signaling concerns

A real-life concern for taking our mechanism to practice is that the monitor may be able to signal her wage to the agent, thus undoing the endogenous asymmetric information that the principal creates. In Section 4 we address this issue by showing that this endogenous adverse selection is robust to settings in which the agent can design a mechanism to extract the monitor's (endogenous) private information.

However, in real-life a monitor may able to signal her wage by taking costly actions that lie outside of our model. For instance, a monitor with a high wage may be able to truthfully signal her level of compensation by buying luxury items (i.e., an expensive watch) that low paid monitors cannot afford. A practical way in which the principal can lessen these concerns is by giving high paid monitors the same *current* wage as low paid monitors, and by collecting their additional wage into a deferred compensation account which can only be accessed by the monitor if she performs well (i.e., if she is not caught lying).

6.3 The impact of extortion

Our model assumes that the monitor always send the message that maximizes her expected payoff in case she doesn't collude with the agent. This assumption implies that the monitor will not get a bribe when the agent takes action $e = 0$: since the monitor's expected wage from telling the truth is higher than her expected wage from lying, in this case the agent doesn't need to bribe the monitor to get a favorable report. We now show how our results extend to settings in which the monitor can extort an agent who took action $e = 0$ by committing to send an untruthful message. The framework we consider is essentially the same as in Section 4. The only difference is that a monitor who makes the offer at the side-contracting stage can commit to send an untruthful message to the principal if the agent rejects her proposal.

In Appendix B.3 we show that, in this setting, the expected payoff an agent with type α gets from taking action $e = 0$ is equal to $-\lambda(1-q)t$. The term $(1-q)t$ represents the bribe that the agent pays when the monitor makes the offer at the side-contracting stage. On the other hand, by the same arguments as in Section 4, the agent's payoff from taking action $e = 1$ and making offer $T \geq 0$ when selected to make a proposal is $\alpha - t + (1-\lambda)F(T/q)((1-q)t - T)$. An agent with type α then takes action $e = 0$ if and only if,

$$\forall T \geq 0, \ \alpha - t + (1-\lambda)F(T/q)((1-q)t - T) \leq -\lambda(1-q)t. \tag{8}$$

Therefore, in this setting the problem of a principal who wants to give incentives to agents with type $\alpha' \leq \alpha$ to take action $e = 0$ is

$$\inf_{F \in \mathcal{F}} \int w dF(w) \text{ s.t. (8)} \tag{9}$$

Proposition 7. *Suppose the monitor can extort the agent when she takes action $e = 0$. Then, it is optimal for the principal to pay the monitor a random wage drawn from the cdf $\hat{F}_\alpha \in \mathcal{F}$ such that:*

$$\hat{F}_\alpha(w) = \min\left\{\frac{t - \alpha - \lambda(1-q)t}{(1-\lambda)((1-q)t - qw)}, 1\right\}. \tag{10}$$

Proof. Appendix B.3 □

The distribution $\hat{F}_\alpha$ in Proposition 7 first order stochastically dominates the distribution F_α in Proposition 1. Therefore, when there is a threat that monitors might extort agents the principal needs to increase wage payments in order to provide incentives.

6.4 An alternative implementation

This paper shows how the principal can improve incentive provision by adding asymmetric information between the monitor and the agent. The way the principal generates this endogenous adverse selection in our model is by randomizing over the monitor's compensation scheme and by informing only the monitor about the outcome of this randomization. This subsection illustrates an alternative way in which a principal can add asymmetric information between the monitor and the agent.

Consider the same setting as in Section 2, with the added feature that the principal can choose the probability q with which she detects false messages. For simplicity, we focus on the case in which the agent has all the bargaining power at the side-contracting stage. For all $q \in [0, 1]$, let $c(q)$ be the cost that the principal has to incur to detect false messages by the monitor with probability q. For instance, the principal may affect the likelihood with which he detects false messages by choosing how much effort to spend checking reports.

Suppose the principal offers the monitor a deterministic efficiency wage $w > 0$, but randomizes over the probability q with which he detects false messages. Suppose further that the principal informs only the monitor about the outcome of this randomization. As before, this randomization adds asymmetric information between the monitor and the agent.

Consider first a setting in which the principal uses a deterministic strategy $q \in (0, 1)$. In this case, an agent who takes action $e = 1$ would have to offer the monitor a transfer $T \geq qw$ to induce her to send a favorable report to the principal. The agent's payoff from taking action $e = 1$ and making offer $T = qw$ would then be $\alpha - qt - qw$. On the other hand, the agent obtains a payoff of 0 from taking action $e = 0$, since in this case the monitor has an incentive to send a truthful message $m = 0$ even without a bribe. It follows that q must be at least as large as $\frac{\alpha}{t+w} \in (0, 1)$ in order to provide incentives to the agent.

Consider next a setting in which that the principal randomizes over the choice of q. For instance, suppose that the principal chooses $q = 0$ with probability $x \in [0, 1]$ and $q = \hat{q} \leq 1$

with probability $1 - x$. As before, the agent obtains a payoff of 0 by taking action $e = 0$. If the agent instead takes action $e = 1$, there are two potentially optimal offers she can make to the monitor: an offer $T = 0$ which the monitor accepts only if $q = 0$, or an offer $T = \hat{q}w$ which the monitor always accepts. The agent's payoff from making offers $T = 0$ and $T = \hat{q}w$ is, respectively, $\alpha - (1 - x)t$ and $\alpha - (1 - x)\hat{q}t - \hat{q}w$. If $x = \frac{t-\alpha}{t}$ and $\hat{q} = \frac{\alpha}{\alpha+w}$, the agent's payoff from taking action $e = 1$ is less than zero regardless of the offer she makes. Moreover, the expected value of q is $x \times 0 + (1 - x) \times \hat{q} = \frac{\alpha}{t}\frac{\alpha}{\alpha+w} < \frac{\alpha}{t+w}$. Therefore, if the cost function $c(q)$ is not too convex, the principal's expected cost of providing incentives to the agent would be strictly lower when he randomizes over q than when he uses a deterministic q.

Appendix

A Extensions

A.1 General wage contracts

Throughout the paper we assume that the principal compensates the monitor with an efficiency wage contract. This appendix shows how our results extend when we allow for general wage contracts $w(m, s)$ in which the monitor's compensation depends on the message m she sends and on the principal's signal s (i.e., on whether the principal detects that the message is false or not). We consider the same model as in the main text. The only difference is that we impose a participation constraint that the agent's payoff cannot be negative.[9]

Suppose the principal wants to induce all agents with type $\alpha' \leq \alpha$ to take action $e = 0$. Since the agent's payoff should be at least zero, agents with type $\alpha' \leq \alpha$ must pay a bribe $T = 0$ to the monitor and get punished with probability 0 whenever they take action $e = 0$: otherwise the agent would get negative payoff if she pays a positive transfer to the monitor and/or if she gets punished with positive probability. This implies that the wage contract $w(m, s)$ must provide incentives to the monitor to report truthfully (even without a transfer) when agent takes action $e = 0$. That is, any wage contract $w(m, s)$ that the principal offers the monitor with positive probability must satisfy $w(0, 0) \geq (1 - q)w(1, 1) + qw(1, 0)$: when the agent takes action $e = 0$ the monitor gets a wage $w(0, 0)$ by reporting truthfully, and she gets an expected wage of $(1 - q)w(1, 1) + qw(1, 0)$ by reporting $m = 1$.

Consider an agent with type α who takes action $e = 1$. By the same arguments as in Section 4, when the monitor is selected to make an offer at the side-contracting stage she will either find it optimal to make an offer $T = (1 - q)t$ that the agent barely accepts, or she

[9] Note that this constraint would not binding when the monitor is compensated with an efficiency wage: in this case, by taking a action $e = 0$, the agent can guarantee herself a payoff of 0 since the monitor will have an incentive to send a truthful message. Therefore, adding this constraint wouldn't change the results in the main text. When we allow for arbitrary wage contracts $w(m, s)$ this constraint rules out wage contracts $w(m, s)$ under which the agent needs to bribe the monitor to get a favorable report after taking action $e = 0$.

will find it optimal to make an offer $T > (1-q)t$ that the agent rejects. In either case, the agent's payoff is equal to $\alpha - t$.

When the agent is selected to make an offer after taking action $e = 1$, the payoff that a monitor with wage contract $w(m, s)$ gets from accepting an offer T is $T + (1-q)w(0,0) + qw(0,1)$: with probability $1-q$ the principal doesn't detect the lie and the monitor gets a wage $w(0,0)$, and with probability q the principal detects the lie and the monitor gets a wage $w(0,1)$. On the other hand, if the monitor rejects the offer she sends message $m = 1$ and gets a payoff of $w(1,1)$. Therefore, the monitor accepts an offer $T \geq 0$ if and only if $T \geq w(1,1) - (1-q)w(0,0) - qw(0,1)$.

Note that it is optimal for the principal to make the wage $w(0,0)$ as low as possible: lowering this wage reduces wage payments, and it also makes the monitor less likely to accept a bribe by an agent who took action $e = 1$. Therefore, for any wage contract $w(m, s)$ that the principal offers the monitor with positive probability, it is optimal to set $w(0,0) = (1-q)w(1,1)+qw(1,0)$. When the wage contract satisfies this property, a monitor accepts an offer T after the agent took action $e = 1$ if and only if $T \geq w(1,1)-(1-q)w(0,0)-qw(0,1) = w(1,1)q(2-q) - q(w(1,0) + (1-q)w(0,1))$. Note that it is optimal for the principal to set $w(1,0) = w(0,1) = 0$ for all wage contracts that he offers the monitor with positive probability, since doing this reduces the monitor's incentive to accept a bribe from an agent who took action $e = 1$. This implies that any wage contract $w(m, s)$ that the principal offers is fully characterized by $w(1,1)$, since $w(0,0) = (1-q)w(1,1)$ and $w(1,0) = w(0,1) = 0$.

Let $F \in \mathcal{F}$ be the cdf of wages $w(1,1)$ that the principal uses. By the arguments in the previous paragraphs, the monitor accepts an offer $T \geq 0$ from an agent who took action $e = 1$ if and only if $T \geq \hat{q}w(1,1)$, where $\hat{q} = q(2-q)$. Therefore, from the agent's perspective, an offer $T \geq 0$ is accepted by the monitor with probability $F(T/\hat{q})$. The agent's expected payoff from taking action $e = 1$ and making offer T when she is selected to make offers is then given by $\lambda(\alpha - t) + (1-\lambda)(\alpha + F(T/\hat{q})(-T - qt) + (1 - F(T/\hat{q}))(-t)) = \alpha - t + (1-\lambda)F(T/\hat{q})((1-q)t - T)$.

Since the agent gets a payoff of 0 by taking action $e = 0$, it follows that she will take action $e = 0$ if and only if,

$$\forall T \geq 0,\ (1-\lambda)F(T/\hat{q})((1-q)t - T) \leq t - \alpha. \tag{A.1}$$

The problem of the principal is to choose a distribution $F \in \mathcal{F}$ that minimizes expected wage payments, subject to the constraint that the agent takes action $e = 0$. Recall that F is the cdf over wages $w(1,1)$. Moreover, if the agent takes action $e = 0$, the monitor will report truthfully and will receive a compensation $w(0,0) = (1-q)w(1,1)$ from the principal. Therefore, the principal's problem is

$$\inf_{F\in\mathcal{F}} \int (1-q)w dF(w) \text{ s.t. } (A.1) \tag{A.2}$$

Following the same argument as in the proof of Proposition 1, the solution to (A.2) is

$$\tilde{F}_\alpha(w) = \min\left\{\frac{t-\alpha}{(1-\lambda)((1-q)t-\hat{q}w)}, 1\right\}.$$

The expected wage payments to the monitor from paying her a random wage contract drawn from the distribution $\tilde{F}_\alpha$ is

$$\int (1-q)wd\tilde{F}_\alpha(w) = \frac{1-q}{2-q}\left(\frac{\alpha - t + (1-q)t(1-\lambda)}{q(1-\lambda)} + \frac{t-\alpha}{q(1-\lambda)}\ln\left(\frac{t-\alpha}{(1-\lambda)(1-q)t}\right)\right).$$

Comparing the expression above to (5), the expected wage payments to the monitor in this setting are a fraction $\frac{1-q}{2-q}$ of the expected wage payments when the principal is restricted to use efficiency wages.

A.2 General bilateral mechanisms

The model in the main text simplifies the side-contracting stage by assuming that either the agent or the monitor make a take-it-or-leave-it offer. In that setting, the monitor's bargaining power vis-a-vis the agent is measured by the likelihood with which she makes an offer. This appendix extends the model in Section 4 by assuming that, at the side-contracting stage, the monitor and the agent write an individual rational and incentive compatible bilateral bargaining mechanism that maximizes the weighted sum of their payoffs.

Let $\mathcal{F}_I \subset \mathcal{F}$ be the set of cdfs whose support takes the form $[\underline{w}, \overline{w}] \subset [0, \infty)$. Suppose that the principal randomizes over the monitor's efficiency wage according to some cdf $F \in \mathcal{F}_I$.[10] If the agent takes action $e = 0$, the monitor finds it weakly optimal to send a truthful message $m = 0$ to the principal even without getting a transfer from the agent. Therefore, the agent gets a payoff of 0 by taking action $e = 0$, regardless of the bilateral bargaining mechanism that the monitor and agent use at the side-contracting stage.

When the agent takes action $e = 1$, the monitor would find it optimal to send a message $m = 1$ to the principal if there is no agreement at the side-contracting stage. The agent would incur a punishment cost of $t > 0$ if the monitor sends this message. On the contrary, the agent would incur an expected punishment cost of qt if the monitor sends message $m = 0$. Sending a false message has a cost of qw for a monitor with wage w, so the total gains from sending message $m = 0$ instead of $m = 1$ are $(1 - q)t - qw$. When the agent takes action $e = 1$, the agent and the monitor bargain over how to split these gains.

To determine how the monitor and the agent split these gains from reaching an agreement, we consider the class of *bilateral bargaining mechanisms* that are individually rational and incentive compatible. By the revelation principle, we can focus on direct revelation mechanisms in which the monitor reports her (endogenous) private information; i.e., her wage. A bilateral bargaining mechanism is characterized by two functions: (i) $P(w)$, the

[10] Restricting attention to cdfs in $\mathcal{F}_I$ allows us to use the envelope formula to calculate the payoffs of the monitor and the agent from any bilateral mechanism.

probability with which the monitor and the agent reach an agreement when the monitor's wage is w; and (ii) $T(w)$, the transfer from the agent to the monitor when the monitor's wage is w. The monitor commits to send message $m = 0$ if there is an agreement. Otherwise, if there is no agreement the monitor sends a truthful message $m = 1$.

Given a cdf over wages $F \in \mathcal{F}_I$, the expected payoff of an agent who took action $e = 1$ from a bilateral bargaining mechanism (P, T) is

$$\begin{aligned} U_A &= \int (\alpha - T(w) + P(w)(-qt) + (1 - P(w))(-t))\, dF(w) \\ &= \int (P(w)(1 - q)t - T(w))\, dF(w) + \alpha - t, \end{aligned}$$

The agent's individual rationality constraint is $U_A \geq \alpha - t$.

On the other hand, the payoff of a monitor with wage w and who reports wage w' is $U_M(w, w') = P(w')(1 - q)w + T(w') + (1 - P(w'))w = T(w') + (1 - P(w')q)w$. By incentive compatibility, it must be that $U_M(w, w) \geq U_M(w, w')$ for all $w' \neq w$. By individual rationality it must be that $U_M(w, w) \geq w$.

Given a distribution $F \in \mathcal{F}_I$, the weighted sum of the agent's and monitor's payoff is

$$\lambda \int \tilde{U}_M(w) dF(w) + (1 - \lambda)U_A, \tag{A.3}$$

where $\lambda \in [0, 1]$ is the weight on the monitor's payoff. For every $F \in \mathcal{F}_I$ and every $\lambda \in [0, 1]$, let $\Gamma(F, \lambda)$ be the set of incentive compatible and individually rational bilateral bargaining mechanisms that maximize (A.3). We assume that, at the side-contracting stage, the monitor and the agent use a bilateral bargaining mechanism in $\Gamma(F, \lambda)$. The parameter λ measures the monitor's bargaining power. Let $\tilde{U}_A^{\alpha}(F, \lambda)$ be the lowest utility that an agent with type α gets under a bargaining mechanism in $\Gamma(F, \lambda)$ if she takes action $e = 1$.

The problem of a principal who wants to incentivize agents with type $\alpha' \leq \alpha$ to take action $e = 0$ is to choose a distribution of wages $F \in \mathcal{F}_I$ that minimizes expected wage payments, subject to the constraint that such agents weakly prefer to take action $e = 0$ than action $e = 1$. That is, the principal's problem is

$$\inf_{F \in \mathcal{F}_I} \int w dF(w) \text{ s.t.} \tilde{U}_A^{\alpha}(F, \lambda) \leq 0 \tag{A.4}$$

Proposition A1. *Suppose that the principal wants to provide incentives to all agents with type below α to take action $e = 0$. If $\lambda \in [0, 1/2)$, it is optimal for the principal to pay the monitor a random wage drawn from the distribution $\bar{F}_{\alpha} \in \mathcal{F}$ such that, for all $w \geq 0$,*

$$\bar{F}_{\alpha}(w) = \min \left\{ \left(\frac{t - \alpha}{(1-q)t - qw} \right)^{\frac{1-2\lambda}{1-\lambda}}, 1 \right\}. \tag{A.5}$$

When $\lambda \in [1/2, 1]$, it is optimal for the principal to pay the monitor a wage of 0 with probability 1.

Proof. See Appendix B.4. □

Proposition A1 generalizes the results in Section 4 to settings in which the monitor and the agent can use an optimal bilateral bargaining mechanism at the side-contracting stage. When $\lambda \in [0, 1/2)$, $\bar{F}_{\alpha}(w)$ is increasing in λ. This implies that it is cheaper for the principal to provide incentives to the agent when the monitor has more bargaining power at the side-contracting stage. This comparative statics was also present in the model in Section 4, and its intuition is the same as in that setting.

For all $\lambda \geq 1/2$, the distribution that solves (A.4) has all its mass at $w = 0$. In these cases, the principal can incentivize the agent to take action $e = 0$ at no cost. The proof of Proposition A1 shows that, for all $\lambda \geq 1/2$, the incentive compatible and individually rational bargaining mechanism that maximizes (A.3) is the mechanism that maximizes the monitor's expected payoff. In this case, the monitor extracts all the benefits that the agent derives from a favorable message after she takes action $e = 1$, so the principal can provide incentives to the agent by paying a wage $w = 0$ to the monitor with probability 1. Finally, note that the the distribution $\bar{F}_\alpha$ is continuous in λ: as $w \geq 0$, $\bar{F}_\alpha(w) \to 1$ as $\lambda \uparrow 1/2$.

A.3 Ex-ante collusion

The model in the main text assumes that the monitor and the agent can collude after the agent takes action $e \in \{0, 1\}$. This appendix studies the role of random incentives in settings in which the monitor and the agent can collude before the agent chooses her action.

We consider a model in which the agent chooses which action $e \in \{0, 1\}$ to take after side-contracting with the monitor, but which is otherwise the identical to the model in the main text. We make two simplifying assumptions: (i) the private benefit α that the agent derives from taking action $e = 1$ is common knowledge and satisfies $\alpha \in (qt, t)$, and (ii) the agent has all the bargaining power at the side-contracting stage. At the side-contracting stage the agent makes a take-it-or-leave-it offer $T \geq 0$ to the monitor. If the monitor accepts the agent's offer, she commits to send a message $m = 0$ to the principal regardless of the agent's action. Otherwise, if the monitor rejects the agent's offer, she sends a truthful message to the principal (since this is the message that maximizes her expected wage payments).

Suppose that the monitor accepts the agent's offer T. In this case, the agent's payoff from taking action $e = 1$ is $\alpha - qt - T$, since she will only get punished when the principal detects that the agent had lied. On the other hand, by taking action $e = 0$ the agent gets a payoff of $-T < \alpha - qt - T$. Therefore, the agent always takes action $e = 1$ after the monitor accepts her offer. On the other hand, the agent always takes action $e = 0$ if the monitor rejects her offer: in this case she obtains a payoff of 0 by taking $e = 0$, while her payoff from taking $e = 1$ is $\alpha - t < 0$.

Consider the best-response of a monitor with efficiency wage w to an offer T by the agent. The monitor's payoff from accepting such an offer is $T + (1 - q)w$, since in this case the agent will take action $e = 1$ and the principal will detect that the monitor had lied with probability q. On the other hand, by rejecting the offer the monitor obtains a wage payment

w. Therefore, the monitor accepts the agent's offer T if and only if $T \geq qw$.

Suppose the principal pays the monitor an efficiency wage drawn from the cdf $F \in \mathcal{F}$. Then, the agent's payoff from making an offer $T \geq 0$ is $F(T/q)(\alpha - qt - T)$: the agent takes action $e = 1$ if the offer is accepted and obtains a payoff of $\alpha - qt - T$; otherwise, the agent takes action $e = 0$ if the offer is rejected and obtains a payoff of 0. The agent's optimal offer is $T^* \in \arg\max F(T/q)(\alpha - qt - T)$.

The principal's problem is

$$\begin{aligned} &\sup_{F \in \mathcal{F}} -F(T^*/q)\beta - \int_0^{T^*/q} (1-q)w dF(w) - \int_{T^*/q}^{\infty} w dF(w) \text{ s.t.} \\ &T^* \in \arg\max F(T/q)(\alpha - qt - T) \end{aligned} \tag{A.6}$$

When $w \leq T^*/q$ the agent successfully bribes the monitor and takes action $e = 1$. In this case, the principal incurs a cost β and pays the monitor an expected wage of $(1-q)w$. When $w > T^*/q$, the monitor rejects the agent's offer, the agent takes action $e = 0$ and the principal pays the monitor her wage w with probability 1.

Suppose first that the principal offers a deterministic wage w to the monitor. If $w < \alpha/q - t$, the agent will find it profitable to offer $T = qw$ and take action $e = 1$. Otherwise, if $w \geq \alpha/q - t$ the agent will find it optimal to take action $e = 0$. Thus, the principal's payoff in this case is $\max\{-\beta, -(\alpha/q - t)\}$: the principal can either pay the monitor a high wage and induce the agent to take action $e = 0$, or can pay the monitor a wage of zero and let the agent take action $e = 1$.

The following proposition characterizes the optimal randomization by the principal.

Proposition A2. *Suppose the agent and the monitor can collude ex-ante. Then, it is optimal for the principal to pay the monitor a random wage drawn from the cdf $F_\alpha^{EA} \in \mathcal{F}$ such that, for all $w \geq 0$,*

$$F^{EA}(w) = \min\left\{ \frac{e^{\frac{-q\beta}{\alpha - qt}}(\alpha - qt)}{\alpha - qt - qw}, 1 \right\}. \tag{A.7}$$

Table A1: Principal's payoff with ex-ante collusion

Parameters	Deterministic wage	Optimal distribution
$\alpha = 2.5, \beta = 20, t = 5, q = 0.1$	-20	-12.6
$\alpha = 2, \beta = 20, t = 5, q = 0.1$	-15	-11
$\alpha = 1.5, \beta = 20, t = 5, q = 0.1$	-10	-8.6

When the principal pays the monitor a wage drawn from F^{EA}, the agent makes an offer $T^* = 0$ to the monitor, and monitor and agent collude with probability $F^{EA}(0) \in (0, 1)$.

Proof. Appendix B.5 □

Proposition A2 shows that, when collusion is ex-ante, the principal finds it optimal to let the monitor and the agent collude a fraction of the time. Intuitively, when collusion is ex-ante the only way in which the principal can deter the agent from taking action $e = 1$ is by always paying the monitor a wage $w = \alpha/q - t$: if the principal pays lower wages with positive probability, the agent will make an offer $T \geq 0$ that a fraction of low paid monitors will accept and will take action $e = 1$ every time she faces a monitor with a sufficiently low wage. If principal pays wages lower than $w = \alpha/q - t$ with positive probability, then the monitor and the agent will collude a fraction of the time. The optimal randomization in Proposition A2 balances the cost β of letting the monitor and the agent collude, and the benefit of paying lower expected wages to the monitor.

The principal's payoff from paying the monitor a wage drawn from the cdf F^{EA} is

$$-F^{EA}(0)\beta - \int w dF^{EA}(w) = -e^{\frac{-q\beta}{\alpha - qt}}\beta - \left(\frac{\alpha}{q} - t\right)\left(1 - e^{\frac{-q\beta}{\alpha - qt}}\left(1 - \frac{-q\beta}{\alpha - qt}\right)\right).$$

Table A1 compares the principal's payoff for different parameter values under the optimal deterministic wage and under the optimal distribution in Proposition A2.[11] As the table shows, the gains from using random incentive contracts can also be substantial in this setting.

[11] The parameters in Table 2 are all such that $\beta > \alpha/q - t$, so the optimal deterministic wage is $\alpha/q - t$.

B Proofs

B.1 Proofs for Section 4

Proof of Proposition 3. Fix a mechanism (P,T). If a monitor with wage w reports that her wage is w', she obtains a payoff $U_M(w,w') = T(w') + P(w')(1-q)w + (1-P(x))w = w(1-qP(w')) + T(w')$: by reporting w', she gets a transfer $T(w')$, an expected wage of $(1-q)w$ if she reaches an agreement with the agent, and a wage of w is there is no agreement. By incentive compatibility, it must be that $U_M(w,w) \geq U_M(w,w')$ for all w'. From these incentive compatibility constraints, it follows that for all w, w',

$$(w-w')(1-qP(w)) \geq U_M(w,w) - U_M(w',w') \geq (w-w')(1-qP(w')).$$

The equation above imply that $P(w)$ is decreasing in w and that $U_M(w,w) = \int_w^{\overline{w}_\alpha} qP(\hat{w})d\hat{w} + w + c$ for some constant c. By individual rationality $U_M(w,w) \geq w$ for all w, since a monitor with wage w can always get a payoff of w by not participating in the mechanism and sending the principal a truthful message. Since $U_M(\overline{w}_\alpha, \overline{w}_\alpha) = \overline{w}_\alpha + c$, it must be that $c \geq 0$.

Since $U_M(w,w) = T(w) + (1-qP(w))w = w + \int_w^{\overline{w}_\alpha} qP(\hat{w})d\hat{w} + c$, it follows that $T(w) = P(w)qw + \int_w^{\overline{w}_\alpha} qP(\hat{w})d\hat{w} + c$. Note that we can set $c = 0$, as this minimizes the transfers to the monitor and hence it maximizes the agent's payoff. Replacing this in (6),

$$\begin{aligned} U_A^\alpha(P,T) &= \alpha - t + (1-\lambda)\left[\int_0^{\overline{w}_\alpha} P(w)((1-q)t - qw)dF_\alpha(w) - \int_0^{\overline{w}_\alpha}\int_w^{\overline{w}_\alpha} qP(\hat{w})d\hat{w}dF_\alpha(w))\right] \\ &= \alpha - t + (1-\lambda)\left[\int_0^{\overline{w}_\alpha} P(w)((1-q)t - qw)dF_\alpha(w) - \int_0^{\overline{w}_\alpha} qP(w)F_\alpha(w)dw\right], \end{aligned}$$

where the second inequality follows from changing the order of integration. Note that the cdf $F_\alpha(w)$ has a jump at 0, and then has a strictly positive density $F'_\alpha(w)$ for all $w \in (0, \overline{w}_\alpha)$.

Therefore, the expression above can be written as

$$U_A^\alpha(P,T) \;=\; \alpha - t + (1-\lambda)\left[F_\alpha(0)(1-q)t + \int_0^{\overline{w}_\alpha} P(w)\left((1-q)t - qw - q\frac{F_\alpha(w)}{F'_\alpha(w)}\right)F'_\alpha(w)dw\right].$$

Using $F_\alpha(w)$ in Proposition 1, one can check that $(1-q)t-qw-q\frac{F_\alpha(w)}{F'_\alpha(w)} = 0$ for all $w \in (0, \overline{w}_\alpha]$. Hence, $U_A^\alpha(P,T) = \alpha - t + (1-\lambda)F_\alpha(0)(1-q)t = 0$ for all mechanisms (P,T), where the second equality follows since $F_\alpha(0) = \frac{t-\alpha}{(1-\lambda)(1-q)t}$. □

Proof of Proposition 5. Note first that the principal can provide incentives to agents with type $\alpha \leq t[1-(1-\lambda)(1-q)]$ at zero cost using a deterministic wage $w = 0$. Therefore, to prove the statement of Proposition 5 we can focus on the case with $\alpha > t[1-(1-\lambda)(1-q)]$.

For each $\alpha > t[1-(1-\lambda)(1-q)]$, let $R(\alpha) = \int w dF_\alpha(w)/(1 - \frac{t-\alpha}{(1-\lambda)(1-q)t})\overline{w}_\alpha$ be the ratio of the cost of providing incentives with the optimal distribution F_α to the cost of providing incentives with the two-wage distribution. To prove Proposition 5, it suffices to show that $R(\alpha) \geq 1/2$ for all for all $\alpha > t[1-(1-\lambda)(1-q)]$.

Using (5), for all $\alpha > t[1-(1-\lambda)(1-q)]$

$$R(\alpha) = \frac{\frac{\alpha-t+(1-q)t(1-\lambda)}{q(1-\lambda)} + \frac{t-\alpha}{q(1-\lambda)}\ln\left(\frac{t-\alpha}{(1-\lambda)(1-q)t}\right)}{(1-\frac{t-\alpha}{(1-\lambda)(1-q)t})\frac{\alpha-t+(1-q)t(1-\lambda)}{q(1-\lambda)}}.$$

Applying L'Hopital's rule twice in the expression above, it follows that $\lim_{\alpha\to t[1-(1-\lambda)(1-q)]} R(\alpha) = 1/2$. Given this, to complete the proof of the proposition it suffices to show that $R(\alpha)$ is increasing in α for all $\alpha > t[1-(1-\lambda)(1-q)]$. From the expression above,

$$R'(\alpha) = -\frac{(1-q)t(1-\lambda)\left[2(\alpha - t(q+\lambda-q\lambda)) + (t(2-q-(1-q)\lambda)-\alpha)\ln\left(\frac{t-\alpha}{(1-q)t(1-\lambda)}\right)\right]}{(\alpha - t(q+\lambda-q\lambda))^3}.$$

To show that $R'(\alpha) \geq 0$ for all $\alpha > t[1-(1-\lambda)(1-q)]$ it suffices to show that $r(\alpha) := 2(\alpha - t(q+\lambda-q\lambda)) + (t(2-q-(1-q)\lambda)-\alpha)\ln\left(\frac{t-\alpha}{(1-q)t(1-\lambda)}\right) \leq 0$ for all such α. Note that $\lim_{\alpha\to t[1-(1-\lambda)(1-q)]} r(\alpha) = 0$ and that $r'(\alpha) = \frac{t(q+\lambda-q\lambda)-\alpha}{t-\alpha} - \ln\left(\frac{t-\alpha}{(1-q)(1-\lambda)t}\right)$. Note further that

$\lim_{\alpha \to t[1-(1-\lambda)(1-q)]} r'(\alpha) = 0$, and that $r''(\alpha) = \frac{t(q+\lambda-q\lambda)-\alpha}{(t-\alpha)^2} < 0$ for all $\alpha > t[1-(1-\lambda)(1-q)]$. It then follows that $r'(\alpha) < 0$ for all $\alpha > t[1-(1-\lambda)(1-q)]$, which, together with $\lim_{\alpha \to t[1-(1-\lambda)(1-q)]} r(\alpha) = 0$, implies that $r(\alpha) < 0$ for all $\alpha > t[1-(1-\lambda)(1-q)]$. Hence, $R(\alpha) > 1/2$ for all $\alpha > t[1-(1-\lambda)(1-q)]$, and so Proposition 5 follows. □

B.2 Analysis for Section 5

We first show the optimal offer of a monitor with wage w is given by (7). The payoff the monitor gets from making offer $(1-q)t_1$ is $(1-q)t_1 + (1-q)w$: all agents accept this offer, the monitor sends message $m = 0$ and the principal detects that the message is false with probability q. The payoff that the monitor gets from making an offer $(1-q)t_2$ is $p((1-q)t_2 + (1-q)w) + (1-p)w$: if the agent's punishment cost is t_2, the agent accepts the offer, the monitor sends message $m = 0$ and the principal detects that the message is false with probability q; if the agent's punishment cost is t_1, the agent rejects the offer and the monitor sends message $m = 1$. The payoff that the monitor gets from making an offer $T > (1-q)t_2$ is w, since all agents reject such an offer. Comparing these three payoffs, it follows that the monitors optimal offer is given by (7). Note that the payoff that an agent with type (α, t_i) gets from accepting an offer $T \leq (1-q)t_i$ after taking action $e = 1$ is $\alpha - qt_i - T$, while the payoff she gets from rejecting the offer is $\alpha - t_i$.

Consider next an agent with type (α, t_i) who took action $e = 1$ and who is selected to make an offer at the side-contracting stage. An offer T by the agent is accepted by the monitor if and only if $T \geq qw$. Therefore, if she makes an offer T, the agent's payoff when she is selected to make offers is $\alpha + F(T/q)(-qt_i - T) + (1 - F(T/q))(-t_i) = \alpha - t_i + F(T/q)((1-q)t_i - T)$.

The expected payoff of an agent with type (α, t_1) from taking action $e = 1$ is therefore $\alpha - t_1 + (1-\lambda)F(T/q)((1-q)t_1 - T)$, since the payoff that such agent gets when the monitor makes offers is $\alpha - t_i$. On the other hand, the expected payoff of an agent with type (α, t_2) from taking action $e = 1$ is $\alpha - t_2 + (1-\lambda)F(T/q)((1-q)t_2 - T) + F(\hat{w}^-)(1-q)(t_2 - t_1)$.

Consider next the problem of a principal who wants to provide incentives to agents with type (α, t_2) (with $\alpha > t_2[1-(1-q)(1-\lambda)]$) to take action $e = 0$ (so agents with type (α', t_2) with $\alpha' \leq \alpha$ would also take action $e = 0$). The principal's problem is

$$\inf_{F \in \mathcal{F}} \int w dF(w) \text{ s.t.} \tag{B.1}$$

$$\forall T \geq 0, (1-\lambda)F(T/q)((1-q)t_2 - T) + F(\hat{w}^-)(1-q)(t_2 - t_1) \leq t_2 - \alpha. \tag{B.2}$$

Proof of Proposition 6. The constraint (B.2) implies that, for all $T \geq q\hat{w}$,

$$F(T/q) \leq \frac{t_2 - \alpha - F(\hat{w}^-)(1-q)(t_2 - t_1)}{((1-q)t_2 - T)} \tag{B.3}$$

Moreover, by the same arguments as in the proof of Proposition 1, it is optimal for the principal to set (B.3) with equality for all $T \geq q\hat{w}$. Applying the change of variable $w = T/q$, the optimal distribution must be such that $F(w) = \frac{t_2-\alpha-F(\hat{w}^-)(1-q)(t_2-t_1)}{((1-q)t_2-qw)}$ for all $w \geq \hat{w}$. Note that, for all $w \geq \hat{w}$, $F(w)$ is strictly decreasing in $F(\hat{w}^-)$. This implies that $\int_{(\hat{w},\infty)} wdF(w)$ is strictly increasing in $F(\hat{w}^-)$.[12]

Let $F^* \in \mathcal{F}$ be such that $F^*(w) = 0$ for all $w < \hat{w}$ and $F^*(w) = \frac{t_2-\alpha}{((1-q)t_2-qw)}$ for all $w \geq \hat{w}$; i.e., F^* is the cdf $F(w) = \frac{t_2-\alpha-F(\hat{w}^-)(1-q)(t_2-t_1)}{((1-q)t_2-qw)}$ with $F(\hat{w}^-) = 0$. The expected wage payments under F^* are $F^*(\hat{w}) \times \hat{w} + \int_{(\hat{w},\infty)} wdF^*(w)$ (where $F^*(\hat{w})$ is the atom that F^* has at $\hat{w}$). Note that any other candidate solution to (B.1) must be such that $F(w) \leq F(\hat{w}^-)$ for all $w \in [0, \hat{w})$ and $F(w) = \frac{t_2-\alpha-F(\hat{w}^-)(1-q)(t_2-t_1)}{((1-q)t_2-qw)}$ for all $w \geq \hat{w}$. Therefore, the expected wage payments under any other candidate solution F must be at least as large as $F(\hat{w}^-) \times 0 + \int_{(\hat{w},\infty)} wdF(w)$. The difference in expected wage payments under these two distributions is $F^*(\hat{w}) \times \hat{w} + [\int_{(\hat{w},\infty)} wdF^*(w) - \int_{(\hat{w},\infty)} wdF(w)]$. By the arguments in the previous paragraph the term in square brackets is strictly negative for all values of t_1, while the first the term goes to zero as $t_1 \downarrow pt_2$ (since $\lim_{t_1 \downarrow pt_2} \hat{w} = 0$). Hence, there exists $\bar{t} > pt_2$

[12] Note that F might have an atom at $\hat{w}$. The integral $\int_{(\hat{w},\infty)} wdF(w)$ denotes the integral in $(\hat{w}, \infty)$, without the (possible) atom at $\hat{w}$.

such that F^* is the optimal distribution for all $t_1 < \bar{t}$. Finally, note that the largest wage in the support of F^* is $\overline{w}_\alpha(t_2)$. □

B.3 Proof of Proposition 7

Consider an agent who takes action $e = 0$, and suppose the monitor is selected to make the offer at the side-contracting stage. Since the monitor can commit to send an untruthful message if the agent rejects the offer, the agent's payoff from rejecting the offer is $-(1-q)t$: in this case, the principal will punish the agent unless he detects that the monitor's message was false (which occurs with probability q). On the other hand, the payoff from accepting the monitor's offer T is $-T$, since in this case the monitor sends a truthful message. Hence, the agent accepts any offer $T \leq (1-q)t$. A monitor always finds it strictly optimal to make an offer $T = (1-q)t$: by making such an offer she obtain a payoff $(1-q)t + w$. Therefore, the agent's payoff from taking action $e = 0$ when the monitor makes the offer is $-(1-q)t$.

Consider next the case in which the agent who takes action $e = 0$ and is selected to make an offer at the side-contracting stage. Given our assumption that the monitor can only commit to send an untruthful message when she makes the offer, in this case the monitor will send a truthful message even if the agent makes an offer $T = 0$, so the agent gets a payoff of zero. The agent's expected payoff from taking action $e = 0$ is then given by $-\lambda(1-q)t$.

On the other hand, by the same arguments in Section 4, the agent's expected payoff from taking action $e = 1$ is $\alpha - t + (1-\lambda)F(T/q)((1-q)t - T)$. Hence, an agent with type α will take action $e = 0$ if and only if, for all $T \geq 0$, $\alpha - t + (1-\lambda)F(T/q)((1-q)t - T) \leq -\lambda(1-q)t$.

Proof of Proposition 7. By the same arguments as in the proof of Proposition 1, if F is solves (9) it must be that the constraint (8) holds with equality for all offers $T \geq 0$ such that $F(T/q) < 1$. Therefore, the solution to (9) satisfies

$$(1-\lambda)F(T/q)((1-q)t - T) = t - \alpha - \lambda(1-q)t \Rightarrow F(T/q) = \frac{t - \alpha - \lambda(1-q)t}{(1-\lambda)((1-q)t - T)}. \quad \text{(B.4)}$$

Applying the change of variable $w = T/q$ in (B.4), it follows that F satisfies (10). □

B.4 Proof of Proposition A1

Fix $F \in \mathcal{F}_I$ with support $[\underline{w}, \overline{w}]$, $\lambda \in [0,1]$ and a mechanism (P,T), and note that

$$\lambda \int_{\underline{w}}^{\overline{w}} \tilde{U}_M(w) dF(w) + (1-\lambda) U_A =$$
$$\int_{\underline{w}}^{\overline{w}} \left[P(w)\left((1-\lambda)(1-q)t - \lambda q w\right) + (2\lambda - 1)T(w) + \lambda w\right] dF(w) + (1-\lambda)(\alpha - t). \quad \text{(B.5)}$$

By the same arguments as in the proof of Proposition 3, any incentive compatible mechanism (P,T) must be such that $P(w)$ is decreasing, $U_M(w,w) = \int_w^{\overline{w}} qP(\hat{w})d\hat{w} + w + c$ for some constant $c \geq 0$, and $T(w) = P(w)qw + \int_w^{\overline{w}} qP(\hat{w})d\hat{w} + c$ (where $\overline{w}$ is the highest wage in the support of F). Replacing this into (B.5), the weighted sum of payoffs is

$$\begin{aligned}
& \int_{\underline{w}}^{\overline{w}} \left[P(w)(1-\lambda)\left((1-q)t - qw\right) + \lambda w\right] dF(w) + (1-\lambda)(\alpha - t) \\
& +(2\lambda - 1)\left(\int_{\underline{w}}^{\overline{w}} \int_w^{\overline{w}} qP(\hat{w}) d\hat{w} dF(w) + c\right). \\
= \; & \int_{\underline{w}}^{\overline{w}} \left[P(w)(1-\lambda)\left((1-q)t - qw\right) + \lambda w\right] dF(w) + (1-\lambda)(\alpha - t) \\
& +(2\lambda - 1)\left(\int_{\underline{w}}^{\overline{w}} qP(w)F(w)dw + c\right), \qquad \text{(B.6)}
\end{aligned}$$

where the second equality follows after changing the order of integration in $\int_{\underline{w}}^{\overline{w}} \int_w^{\overline{w}} qP(\hat{w})d\hat{w}dF(w)$.

Lemma B1. *For all $\lambda \in [0, 1/2)$, the mechanism (P,T) that maximizes (B.6) is such that $P(w) = \mathbf{1}_{\{w \leq w^*\}}$ for some w^* and $T(w) = P(w)qw + \int_w^{\overline{w}} qP(\hat{w})d\hat{w}$.*

Proof. We first show that, when looking at mechanisms that maximizes (B.6), we can restrict attention to mechanisms with the property that $P(w)$ only takes values 0 or 1. To see this, suppose there exists a interval V such that $P(w) \in (0,1)$ for all $w \in V$, and let $H = \int_V (1-\lambda)[((1-q)t - qw) + \lambda w]dF(w) + (2\lambda - 1)\int_V qF(w)dw$. If $H \geq 0$, then increasing

$P(w)$ over this interval (subject to the constraint that P is decreasing) will make (B.6) larger. If $H < 0$, then decreasing $P(w)$ over this interval (subject to the constraint that P is decreasing) will also make (B.6) larger. Such improvements are exhausted when $P(w)$ only takes values 0 and 1. Since P must be decreasing, if P only takes values on 0 and 1 there must exist a wage w^* such that $P(w) = \mathbf{1}_{\{w \leq w^*\}}$. Finally, note that when $\lambda \in [0, 1/2)$, (B.6) is maximized by setting $c = 0$. Hence, $T(w) = P(w)qw + \int_w^{\overline{w}} qP(\hat{w})d\hat{w}$. □

Lemma B2. *For all $\lambda \in [0, 1/2)$, the solution to problem (A.4) is given by (A.5).*

Proof. By Lemma B1, for all $\lambda \in [0, 1/2)$ the monitor and the agent use a mechanism (P, T) such that $P(w) = \mathbf{1}_{\{w \leq w^*\}}$. Using this, (B.6) can be written as

$$(1-\lambda)[F(w^*)(1-q)t - \int_0^{w^*} qwdF(w) + \alpha - t] + \lambda \int wdF(w) + (2\lambda - 1)\int_0^{w^*} qF(w)dw.$$

Since (P, T) maximizes the weighted sum of payoffs, for all $\hat{w} \neq w^*$ it must be that

$$\begin{aligned} & (1-\lambda)[F(w^*)(1-q)t - \int_0^{w^*} qwdF(w)] + (2\lambda - 1)\int_0^{w^*} qF(w)dw \\ \geq \; & (1-\lambda)[F(\hat{w})(1-q)t - \int_0^{\hat{w}} qwdF(w)] + (2\lambda - 1)\int_0^{\hat{w}} qF(w)dw. \end{aligned} \tag{B.7}$$

Otherwise, if (B.7) did not hold for some $\hat{w} \neq w^*$, the weighted sum of payoffs would be strictly larger under mechanism $(\hat{P}, \hat{T})$ with $\hat{P}(w) = \mathbf{1}_{\{w \leq \hat{w}\}}$.

The agent's payoff from this mechanism is equal to $\alpha - t + \int_0^{w^*} [P(w)(1-q)t - T(w)]dF(w)$. Recall that $T(w) = P(w)qw + \int_w^{\overline{w}} qP(\hat{w})d\hat{w}$. Since $P(w) = \mathbf{1}_{\{w \leq w^*\}}$, it follows that $T(w) = qw^*$ for all $w \leq w^*$, and $T(w) = 0$ for all $w > w^*$. Hence, the agent's payoff from this mechanism is $\alpha - t + F(w^*)[(1-q)t - qw^*]$.

Consider next the principal's problem, who has to choose a distribution $F \in \mathcal{F}_I$ to minimize expected wage payments subject to the constraint that the agent finds it optimal to take action $e = 0$. If F is the optimal distribution, then equation (B.7) must hold with equality for all $\hat{w} \neq w^*$ such that $F(\hat{w}) < 1$. To see this, suppose by contradiction that there exists w' with $F(w') < 1$ such that (B.7) holds with strict inequality at $\hat{w} = w'$. Let $\tilde{F} \in \mathcal{F}_I$ be such that, for all $\hat{w} \in [0, \tilde{w}]$,

$$\begin{aligned} & (1-\lambda)[\tilde{F}(\hat{w})(1-q)t - \int_0^{\hat{w}} qwd\tilde{F}(w)] + (2\lambda - 1)\int_0^{\hat{w}} q\tilde{F}(w)dw \\ = \ & (1-\lambda)[F(w^*)(1-q)t - \int_0^{w^*} qwdF(w)] + (2\lambda - 1)\int_0^{w^*} qF(w)dw. \end{aligned}$$

where $\tilde{w} := \inf\{w : \tilde{F}(w) = 1\}$ (our arguments below show that such a cdf $\tilde{F}$ exists). Note that $\tilde{F}(s) \geq F(s)$ for all s, with strict inequality for some s. This implies that expected wages under $\tilde{F}$ are lower than under F. Moreover, note that mechanism (P, T) with $P(w) = \mathbf{1}_{\{w \leq w^*\}}$ and $T(w) = qw^*$ is still also an optimal mechanism under $\tilde{F}$, and that the agent's payoff from taking action $e = 1$ under this mechanism when the distribution of wages is $\tilde{F}$ is $\tilde{F}(w^*)((1-q)t - qw^*) + \alpha - t = F(w^*)((1-q)t - qw^*) + \alpha - t \leq 0$. Therefore, the agent would still have an incentive to take action $e = 0$ under the distribution $\tilde{F}$, which contradicts the assumption that F solves (A.4). Hence, (B.7) must hold with equality for all $\hat{w}$ such that $F(\hat{w}) < 1$: the right-hand side (B.7) must be equal to a constant for all such $\hat{w}$.

Note that F is differentiable almost everywhere (being a monotone function). Differentiating the right-hand side of (B.7) with respect to $\hat{w}$ we get that, almost everywhere,

$$F'(\hat{w})(1-\lambda)[(1-q)t - q\hat{w}] + qF(\hat{w})(2\lambda - 1) = 0, \tag{B.8}$$

where the equality follows since the right-hand side of (B.7) is constant. The solution to (B.8) is $F(w) = C\left((1-q)t - qw\right)^{-\frac{1-2\lambda}{1-\lambda}}$ for some constant C. By construction, $F(w)$ is such that $(1-\lambda)[F(\hat{w})(1-q)t - \int_0^{\hat{w}} qwdF(w)] + (2\lambda - 1)\int_0^{\hat{w}} qF(w)dw$ is constant for all $\hat{w}$.

We now determine the value C. For any $\hat{w}$ in the support of F, let $(P_{\hat{w}}, T_{\hat{w}})$ be the mechanism with $P_{\hat{w}}(w) = \mathbf{1}_{\{w \leq \hat{w}\}}$ and $T_{\hat{w}}(w) = q\hat{w}$ for all $w \in [0, \hat{w}]$. Since (B.7) holds with equality for all $\hat{w}$ such that $F(\hat{w}) < 1$, all such mechanisms maximize (B.5); that is, all such mechanisms are in $\Gamma(F, \lambda)$. Recall that $\tilde{U}^{\alpha}_A(F, \lambda)$ is the lowest utility that the agent gets under a mechanism in $\Gamma(F,\lambda)$. The agent's utility under mechanism $(P_{\hat{w}}, T_{\hat{w}})$ is $u(\hat{w}) := F(\hat{w})((1-q)t-q\hat{w})+\alpha-t$. Note that $u'(\hat{w}) = F'(\hat{w})((1-q)t-q\hat{w})-qF(\hat{w}) = qF(\hat{w})[\frac{1-2\lambda}{1-\lambda}-1] \leq 0$, where the second equality follows since F satisfies (B.8). Therefore, the lowest utility that the agent gets under a mechanism in $\Gamma(F,\lambda)$ is $u(\overline{w}) := ((1-q)t-q\overline{w})+\alpha-t$, where $\overline{w}$ is the highest wage in the support of F. Since $u(\overline{w}) \leq 0$, it must be that $\overline{w} \geq \alpha/q-t$. To minimize expected wages, it is optimal to set $\overline{w} = \alpha/q-t$. This implies that $F(\alpha/q-t) = C(t-\alpha)^{-\frac{1-2\lambda}{1-\lambda}} = 1$, so $C = (t-\alpha)^{\frac{1-2\lambda}{1-\lambda}}$. Hence, $F(w)$ is given by (A.5). □

Lemma B3. *For all $\lambda \in [1/2,1]$, the solution to problem (A.4) is $F(w) = 1$ for all $w \geq 0$.*

Proof. Note that (B.5) is increasing in $T(w)$ when $\lambda \geq 1/2$. Therefore, in this case the mechanism (P,T) that maximizes (B.5) must have $T(w)$ as large as possible, subject the agent's IR constraint; that is, subject to $\alpha - t + \int[P(w)(1-q)t - T(w)]dF(w) \geq \alpha - t$. The maximum is then achieved by setting $\int[P(w)(1-q)t - T(w)]dF(w) = 0$.[13] But this implies that, when $\lambda \geq 1/2$, the agent's payoff from action $e = 1$ under a mechanism that maximizes (B.5) is $\alpha - t < 0$, regardless of the distribution of wages F. Therefore, when $\lambda \in [1/2,1]$ the agent has an incentive to take action $e = 0$ even when F has all its mass at $w = 0$. □

Proof of Proposition A1. Follows from Lemmas B2 and B3. □

B.5 Proof of Proposition A2

Proof of Proposition A2. Let $T^* \geq 0$ be the optimal offer of the agent; that is, for all $T \geq 0$,

$$F(T^*/q)[\alpha - qt - T^*] \geq F(T/q)[\alpha - qt - T]. \tag{B.9}$$

Note that, if $F \in \mathcal{F}$ is the solution to problem (A.6), then (B.9) must hold with equality for all T such that $F(T/q) < 1$. To see this, suppose by contradiction that there exists T with $F(T/q) < 1$ such that $F(T^*/q)[\alpha - qt - T^*] > F(T/q)[\alpha - qt - T]$. Let $\tilde{F} \in \mathcal{F}$ be such

[13] Since $T(w) = P(w)qw + \int_w^{\overline{w}} qP(\hat{w})d\hat{w} + c$, this equality can always be achieved by adjusting c.

that $\tilde{F}(T/q)[\alpha - qt - T] = F(T^*/q)[\alpha - qt - T^*]$ for all T such that $\tilde{F}(T/q) < 1$. Note that $T^* \in \arg\max_T \tilde{F}(T/q)[\alpha - qt - T]$, and that $\tilde{F}(T^*/q) = F(T^*/q)$. Hence, the probability that the agent takes action $e = 1$ is the same under both $\tilde{F}$ and F. Moreover, expected wage payments are lower under $\tilde{F}$ than under F, so the principal's payoff is larger under $\tilde{F}$ than under F. This contradicts the assumption that F solves (A.6), so F must be such that (B.9) holds with equality for all T with $F(T/q) < 1$. This implies that $F(T/q) = \frac{F(T^*/q)(\alpha - qt - T^*)}{\alpha - qt - T}$ for all T with $F(T/q) < 1$.

Note next that it is optimal for the principal to choose F such that $T^* = 0$. To see this, suppose F is such that $T^* > 0$. Let $\hat{F}$ be such that $\hat{F}(0) = F(0)$, and that $\hat{F}(T/q) = \frac{\hat{F}(0)(\alpha - qt)}{\alpha - qt - T}$ for all T such that $\hat{F}(T/q) < 1$. Note that the probability that the agent takes action $e = 1$ is the same under $\hat{F}$ than under F. Moreover, $\hat{F}(T/q) > F(T/q)$ for all T such $\hat{F}(T/q) < 1$, so expected wage payments are lower under $\hat{F}$ than under F. Therefore, it is optimal for the principal to choose F such that $T^* = 0$.

Using the change of variable $w = T/q$, the two paragraphs above imply that the solution to (A.6) is of the form

$$F(w) = \min\left\{\frac{F(0)(\alpha - qt)}{\alpha - qt - qw}, 1\right\},$$

with $F(0) \in [0, 1]$. The principal's payoff from using this distribution of wages is

$$-F(0)\beta - \int w dF(w) = -F(0)\beta - \left(\frac{\alpha}{q} - t\right)(1 - F(0) + F(0)\ln F(0)).$$

Taking first order conditions and noting that the expression above is strictly concave in $F(0)$, it follows that it is optimal for the principal to choose $F(0) = e^{\frac{-q\beta}{\alpha - qt}}$. □

References

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

Calzolari, G. and A. Pavan (2006a): "Monopoly with Resale," *Rand Journal of Economics*, 37, 362–375.

——— (2006b): "On the Optimality of Privacy in Sequential Contracting," *Journal of Economic Theory*, 130, 168–204.

Celik, G. (2009): "Mechanism Design with Collusive Supervision," *Journal of Economic Theory*, 144, 69–75.

Chassang, S. and G. Padró i Miquel (2013): "Corruption, Intimidation and Whistleblowing: A Theory of Inference from Unveriable Reports," *Unpublished manuscript.*

Che, Y.-K. and J. Kim (2006): "Robustly Collusion-Proof Implementation," *Econometrica*, 74, 1063–1107.

Ederer, F., R. Holden, and M. Meyer (2013): "Gaming and Strategic Ambiguity in Incentive Provision," *Unpublished manuscript.*

Eeckhout, J., N. Persico, and P. E. Todd (2010): "A Theory of Optimal Random Crackdowns," *American Economic Review*, 100, 1104–1135.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort (2003): "Collusion, Delegation and Supervision with Soft Information," *Review of Economic Studies*, 70, 253–279.

Felli, L. and J. M. Villa-Boas (2000): "Renegotiation and Collusion in Organizations," *Journal of Economics & Management Strategy*, 9, 453–483.

Fudenberg, D. and J. Tirole (1990): "Moral Hazard and Renegotiation in Agency Contracts," *Econometrica*, 58, 1279–1319.

Jehiel, P. (2012): "On Transparency in Organizations," *Unpublished manuscript.*

Laffont, J.-J. and D. Martimort (1997): "Collusion Under Asymmetric Information," *Econometrica*, 65, 875–911.

——— (2000): "Mechanism Design with Collusion and Correlation," *Econometrica*, 68, 309–342.

Lazear, E. P. (2006): "Speeding, Terrorism, and Teaching to the Test," *Quarterly Journal of Economics*, 121, 1029–1061.

Ma, C. (1991): "Adverse Selection in Dynamic Moral Hazard," *Quarterly Journal of Economics*, 106, 255–275.

Mookherjee, D. and M. Tsumagari (2004): "The Organization of Supplier Networks: Effects of Delegation and Intermediation," *Econometrica*, 72.

Rahman, D. (2012): "But Who Will Monitor the Monitor?" *American Economic Review*, 102, 2767–2797.

Rahman, D. and I. Obara (2010): "Mediated Partnerships," *Econometrica*, 78.

Strausz, R. (2006): "Deterministic versus Stochastic Mechanisms in Principal–agent Models," *Journal of Economic Theory*, 128, 306–314.

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