Making Collusion Hard:

Asymmetric Information as a Counter-Corruption Measure[*]

Juan Ortner Sylvain Chassang[†]

Boston University Princeton University

May 31, 2014

Abstract

We study the problem of a principal who relies on the reports of a monitor to provide incentives to an agent. We allow for collusion, so that the agent and monitor can side-contract on what report to send. We show that the principal can benefit from creating endogenous asymmetric information between the agent and the monitor, thereby making side-contracting more difficult. Specifically, it may be optimal to randomize the incentives given to the monitor, and let the magnitude of her incentives serve as her private information vis à vis the agent.

Plausible numerical computations in simple environments suggest that the potential efficiency gains from random incentives can be large. However, in general, the optimality of random incentives will depend on patterns of pre-existing asymmetric information: it is not always effective to add new sources of asymmetric information. We solve for both the Bayesian and max-min optimal policies, as well as provide an experiment-ready framework for prior-free policy evaluation. We show that even though monitors' reports do not provide a reliable measure of actual corruption, it is possible to evaluate local policy changes using only unverified report data.

KEYWORDS: corruption, monitoring, collusion, endogenous asymmetric information, random incentives, bargaining failure, prior-free policy evaluation, structural experiment design.

[*] We are especially thankful to Gerard Padró i Miquel, for many helpful and inspiring discussions. We are indebted to Bart Lipman, Dilip Mookherjee, Andy Newman and to seminar participants at Iowa State, McGill, and the Washington University in Saint Louis for feedback. Chassang gratefully acknowledges funding from the Alfred P. Sloan Foundation, as well as from the National Science Foundation under grant SES-1156154.

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

1 Introduction

This paper explores the idea that the cost of collusion in organizations can be reduced by introducing asymmetric information frictions between the colluding parties. We consider a principal who relies on the reports of a monitor to provide incentives to an agent, and allow the agent and the monitor to side-contract on what report to send. We show that the principal can reduce the cost of incentive provision by randomizing the incentives given to the monitor, and letting the magnitude of those incentives serve as the monitor's private information vis à vis the agent. The optimality of such random incentives depends on patterns of pre-existing asymmetric information, but the efficiency gains are large in plausible settings. We characterize the max-min and Bayesian optimal policies, as well as provide a framework for prior-free policy evaluation.

We study a game between three players — a principal, an agent, and a monitor — in which the agent takes a corruption decision $c \in \{0, 1\}$, where corruption $c = 1$ gives her a private benefit at a cost to the principal. The behavior of the agent is not observed by the principal, but is observed by the monitor, who sends a report $m \in \{0, 1\}$. We think of this report as hard evidence permitting prosecution: report $m = 1$ triggers an exogenous judicial process which imposes a cost k on corrupt agents; report $m = 0$ (which involves suppression of evidence whenever $c = 1$) triggers no such punishment. Finally, although the principal cannot observe the agent's behavior, she can detect misreporting $m \neq c$ with probability q. The monitor is compensated according to a fixed wage w and is fired in the event that the principal detects misreporting. The only policy control available to the principal are the incentives for truth-telling she provides to the agent, captured by the product qw of the likelihood of detection and possible lost wages.

We allow for collusion between the agent and the monitor at the reporting stage (i.e. after the corruption decision is taken). In particular, the monitor can destroy evidence (i.e. report message $m = 0$) against a corrupt agent in exchange for a bribe. We think of the destruction of evidence as happening in front of the agent, so that there is no moral-hazard between the agent and the monitor, and collusion boils down to a bilateral trading problem. Since we know from Myerson and Satterthwaite (1983) that asymmetric information may prevent efficient trade, we study the extent to which the principal can reduce the cost of incentive provision by creating endogenous asymmetric information between the agent and the monitor.

Our model fits a broad class of environments in which an uninformed principal is concerned about collusion between her monitor and the agents the monitor is supposed to inform about. This includes many of the settings that have been brought up in the empirical literature on corruption, for instance collusion between polluting firms and environmental inspectors (Duflo et al., 2013), tax-evaders and customs officers (Fisman and Wei, 2004), public works contractors and local officials (Olken, 2007), and so on. In these settings the principal cannot efficiently monitor agents directly, but may realistically be able to detect misreporting by scrutinizing accounts, performing random rechecks in person or obtaining tips from informed parties (see Chassang and Padró i Miquel (2013) for work on endogenous reporting). Alternatively, the principal may be able to detect misreporting if corruption has delayed but observable consequences, such as environmental pollution, public infrastructure failures, media scandals, and so on.

Our analysis emphasizes three sets of results. The first is that although deterministic incentive schemes are efficient in the absence of collusion, they can become excessively expensive once collusion is allowed. Efficient contracting between the agent and the monitor forces the principal to raise the monitor's wage to the point where the agent and the monitor's joint surplus from misreporting becomes negative. By using random incentives, the principal can reduce the rents of a corrupt agent, which lowers the cost of incentive provision. We make this point using a simple example without pre-existing asymmetric information. In this case, the cost-savings from using random rather than deterministic incentives are large, reaching 50% under plausible parameter specifications.

Our second set of results extends the analysis to environments with pre-existing asymmetric information. In addition to the incentives provided by the principal, the monitor experiences an exogenous privately observed idiosyncratic cost $\eta \geq 0$ for accepting a bribe. We show that the optimality of using random incentives depends on the convexity or concavity of the c.d.f. F_η of idiosyncratic costs η, and characterize the Bayesian optimal wage schedule. Recognizing that the principal may not have well-formed beliefs over the distribution F_η, we also characterize the max-min optimal wage distribution, and show that it coincides with that derived in environments with no pre-existing private information.

Finally, with implementation in mind, we study the possibility of policy evaluation using reporting data from hypothetical randomized controlled experiments on a large population of agent-monitor pairs. We first show that aggregate reports of corruption across different incentive schemes do not allow for reliable policy evaluation. Indeed, reports of corruption depend on both underlying corruption rates, and the monitors' decision to report corruption or not. As a result, it is possible that a new incentive scheme decreases aggregate reports of corruption, while in fact increasing underlying corruption rates. Surprisingly, we are able to show that it is possible to perform prior-free local policy evaluations using conditional report data, i.e. average reports of corruption conditional on incentives. Somewhat counter-intuitively, a local policy change improves on a reference incentive scheme if it is associated with more reports of corruption.

This paper is most closely related to Chassang and Padró i Miquel (2013) who also consider a game between a principal, an agent, and a monitor in which the agent and the monitor may collude. Both papers explore the idea that collusion may be addressed by exploiting informational frictions that make side-contracting difficult. This paper focuses on asymmetric information while Chassang and Padró i Miquel (2013) focus on moral hazard. They study a model in which reports are non-contractible, so that the monitor is subject to moral hazard. The agent can incentivize her preferred report by committing to a retaliation strategy which can depend on observables. Chassang and Padró i Miquel (2013) show that it is important for the principal to garble the information content of the monitor's reports to limit the effectiveness of incentive provision by the agent. In a spirit similar to our local policy evaluation results, Chassang and Padró i Miquel (2013) also offer a framework for prior-free inference from unverifiable reports.

On the applied side, this paper relates to and hopes to usefully complement the growing empirical literature on corruption. We address two aspects of the problem which have been emphasized in the literature, for instance in the recent survey by Olken and Pande (2012).[1] The first is that the effectiveness of counter-corruption schemes may be very different over the short-run and the long run: over time, agents will find ways to game the system and undermine the monitoring structures in charge of evaluating them. We explicitly take into account the possibility of collusion between agents and monitors and propose novel ways to reduce the costs it imposes on organizations. A second difficulty brought up by Olken and Pande (2012) is that reports of corruption do not provide a reliable measure of underlying corruption. Because reports of corruption depend both on underlying corruption, and on endogenous decisions from monitors to report this corruption or not, reported corruption may decrease while actual corruption increases, and inversely. We address this by providing a framework for prior-free policy evaluation which exploits our structural model to back-out measures of underlying corruption using only reporting data. This connects our work to a small set of papers on structural experiment design (see for instance Karlan and Zinman (2009), Ashraf et al. (2010), Chassang et al. (2012), Chassang and Padró i Miquel (2013), Berry et al. (2012)) that takes guidance from structural models to design experiments whose outcome measures can be used to infer unobservable parameters of interest.

On the theory side, our work fits in the literature on collusion in mechanism design developed by Tirole (1986) and Laffont and Martimort (1997, 2000).[2] The main modelling

[1] For recent work on the measurement of corruption, see Bertrand et al. (2007), and Olken (2007). See also the surveys by Banerjee et al. (2013) and Zitzewitz (2012).

[2] See also Baliga and Sjöström (1998), Felli and Villa-Boas (2000), Faure-Grimaud et al. (2003), Mookherjee and Tsumagari (2004), Che and Kim (2006) or Celik (2009).

difference between much of this literature and our work is that we endogenize the difficulty of contracting between the agent and the monitor, and do not assume that the incentive structure is common-knowledge.[3] We also emphasize a direct mechanism design approach in which the policy instruments and information available to the principal are realistically limited. Finally, we explore the question of robust inference which is new to this literature.

Other work has underlined the usefulness of random incentives for reasons unrelated to collusion. In Becker and Stigler (1974) random checks are an optimal response to non-convex monitoring costs. More recently, in work on police crackdowns, Eeckhout et al. (2010) show that in the presence of budget constraints, it may be optimal to provide high powered incentives to a fraction of a population of agents rather than weak incentives to the entire population. In addition Myerson (1986) and more recently Rahman (2012) emphasize the role of random messaging and random incentives in mechanisms, in particular in settings where the principal needs to disentangle the behavior of different parties.[4]

Finally, our results on prior-free policy evaluation relate the paper to a growing applied theory literature which studies contract design from the perspective of a principal who does not necessarily have a single Bayesian prior over the environment, but rather entertains a set of priors consistent with a few moment restrictions that she can impose based on subjective assessments, or objective data. Examples include Hurwicz and Shapiro (1978), Hartline and Roughgarden (2008), Chassang (2013), Frankel (2014), Chassang and Padró i Miquel (2013), Madarász and Prat (2014), Prat (2014) or Brooks (2014).

The paper is organized as follows. Section 2 introduces our framework in the context of a simple example with no pre-existing private information, and delineates the economic forces

[3] In related work, Baliga and Sjöström (1998) consider a setting in which the agent has no resources of her own, so that any promised payment to the monitor must come from the wage she obtains from the principal. Baliga and Sjöström (1998) show that by randomizing over the agents' wages the principal undermines the agent's ability to commit to transfers.

Also related, although not in the context of collusion, Calzolari and Pavan (2006a,b) which show that a monopolist may benefit from selling to different types of buyers with different probabilities to increase the buyers' ability to extract revenue on a secondary market.

[4] Lazear (2006), Strausz (2006), Jehiel (2012), Rahman and Obara (2010) and Ederer et al. (2013) also emphasize the usefulness of random incentives in organizations.

that make random incentives useful. Section 3 extends the analysis to environments with pre-existing asymmetric information, shows that additional asymmetric information need not always be optimal, and solves for both the max-min, and Bayesian optimal policies. Section 4 takes seriously the possibility of implementing random incentive schemes in the field, and offers an experiment-ready framework for policy evaluation using only unverified report data.

2 A Simple Example

2.1 Framework

Players, actions, and payoffs. We consider a game with three players: a principal, an agent and a monitor. The agent takes a corruption decision $c \in \{0, 1\}$ where corruption $c = 1$ gives the agent a benefit $\pi_A > 0$, and comes at a cost $\pi_P < 0$ to the principal. The agent's action is not directly observable to the principal, but is observed by a monitor who chooses to make a report $m \in \{0, 1\}$ to the principal. Report $m = 1$ triggers an exogenous judiciary process that imposes an expected cost $k > \pi_A$ on corrupt agents and (for simplicity) a cost equal to 0 on non-corrupt agents.

Reports by the monitor are scrutinized by the principal, so that false reports $m \neq c$ are detected with probability $q \in (0, 1)$. The monitor is paid according to a fixed wage contract with wage w, but gets fired in the event that the principal finds evidence of misreporting. The monitor is protected by limited liability and cannot be punished beyond the loss of wages. As part of a possible side-contract the agent can make transfers $\tau \geq 0$ to the monitor. Altogether, expected payoffs u_P, u_A, and u_M respectively accruing to the principal, the agent, and the monitor take the form:

$$\begin{aligned}
u_P &= \pi_P \times c \quad -\gamma_w \times w - \gamma_q \times q \\
u_A &= \pi_A \times c \qquad\qquad\qquad\qquad\quad -k \times c \times m \qquad -\tau \\
u_M &= w \qquad\qquad\qquad\qquad\qquad\qquad -q \times w \times \mathbf{1}_{m \neq c} \quad +\tau,
\end{aligned}$$

where γ_w denotes the efficiency cost of raising wages and γ_q captures the principal's cost of attention. We assume for now that parameters π_A, k, and q are known to the principal.

Note that the monitor's incentives for truthful reporting are captured by the expected loss from misreporting qw. For ease of exposition and consistency with the literature, we think of the distribution of wages w as the principal's policy variable. However, we want to highlight that wages w and scrutiny q enter payoffs in symmetric ways, so that our analysis applies without change if scrutiny q is the relevant policy instrument. As we discuss in Section 5, when giving similar monitors different wages raises fairness concerns, scrutiny q may be the more appropriate choice variable.

Timing and Commitment. Our analysis contrasts the effectiveness of incentive schemes under *collusion* and *no-collusion*. The timing of actions is as follows:

1. the principal commits to a distribution of wages w with c.d.f. F_w, and draws a random wage w for the monitor, which is observed by the monitor but not by the agent;

2. the agent makes a corruption decision $c \in \{0, 1\}$;

3. under *collusion*, the agent makes the monitor a take-it-or-leave-it bribe offer τ in exchange for sending message $m = 0$, which the monitor accepts or rejects — we assume perfect commitment so that whenever the monitor accepts the bribe, she does send message $m = 0$; under *no-collusion* nothing occurs;

4. under *no-collusion* or, under *collusion* if there was no agreement in the previous stage, the monitor sends the message m maximizing her final payoff.

Note that we assume that at the collusion stage, if it occurs, the agent has all the bargaining power. We consider more general bargaining structures in Section 3 and in Appendix A. The following observation is useful.

Fact 1. *Under* collusion*, the monitor will accept a bribe τ from a corrupt agent if and only if $\tau > qw$.*[5] *In equilibrium, the agent never offers a bribe $\tau > \pi_A$.*

Under no-collusion*, or if the monitor rejects the agent's offer, the monitor's optimal continuation strategy is to send truthful reports $m = c$.*

It follows from Fact 1 that the expected payoff of a corrupt agent under collusion is $\pi_A - k + \max_\tau (k - \tau)\text{prob}(qw < \tau)$.

We think of non-collusive and collusive environments as respectively capturing short-run and long-run patterns of organizational behavior. In the short run, the agent may take the monitors' behavior as given, and not explore with bribery. In the long run however, as the agent explores the different strategies available to her, she will learn that monitors respond favorably to bribes.

2.2 The value of endogenous asymmetric information

Deterministic wages. We begin by computing the expected cost $\mathbb{E}_{F_w}[w]$ of keeping the agent non-corrupt when the principal can use only deterministic wages.

Fact 2 (collusion and the cost of incentives)**.** *Assume that the principal uses only deterministic wages. Under* no-collusion *the principal can induce the agent to be non-corrupt at 0 cost.*

Under collusion*, the minimum cost of wages needed to induce the agent to be non-corrupt is equal to $\frac{\pi_A}{q}$.*

[5] By convention, we assume that the monitor rejects the agent's offer whenever she is indifferent between accepting and rejecting a bribe.

While deterministic incentive schemes work well under no-collusion, their effectiveness is significantly limited whenever collusion is a possibility. Note that this remains true if several monitors are used and their messages are cross-checked in the spirit of Maskin (1999). As we show in Appendix A, absent asymmetric information, the cost of bribing two monitors is equal to the cost of bribing a single monitor with twice the incentives.

We now show that by randomizing wage w the principal reduces the efficiency of side-contracting between the agent and the monitor, and hence reduces the cost of incentive provision. We solve for the optimal wage distribution F_w^{bmk} which will serve as a useful benchmark in later sections.

Proposition 1 (optimal incentives under collusion)**.** *Under* collusion *it is optimal for the principal to use random wages. The cost-minimizing wage distribution F_w^{bmk} that induces the agent not to be corrupt is described by*

$$\forall w \in [0, \pi_A/q], \quad F_w^{\mathsf{bmk}}(w) = \frac{k - \pi_A}{k - qw}. \tag{1}$$

The corresponding cost of wages $W^{\mathsf{bmk}}(\pi_A) \equiv \mathbb{E}_{F_w^{\mathsf{bmk}}}[w]$ is

$$W^{\mathsf{bmk}}(\pi_A) = \frac{\pi_A}{q}\left[1 - \frac{k-\pi_A}{\pi_A}\log\left(1 + \frac{\pi_A}{k-\pi_A}\right)\right] = \frac{\pi_A}{q} \times \frac{\pi_A}{k} - o\left(\frac{1}{k}\right). \tag{2}$$

The proof of Proposition 1 is instructive.

Proof. A wage distribution F induces the agent to be non-corrupt if and only if, for every bribe offer $\tau \in [0, \pi_A]$, $\pi_A - k + (k - \tau)\text{prob}(\tau > qw) \leq 0$, or equivalently, if and only if, for every $\tau \in [0, \pi_A]$, $F\left(\frac{\tau}{q}\right) \leq \frac{k-\pi_A}{k-\tau}$. Using the change in variable $w = \frac{\tau}{q}$, we obtain that wage distribution F induces the agent to be non-corrupt if and only if,

$$\forall w \in [0, \pi_A/q], \quad F(w) \leq \frac{k - \pi_A}{k - qw}. \tag{3}$$

By first-order stochastic dominance, it follows that in order to minimize expected wages, the optimal distribution must satisfy (3) with equality. This implies that the optimal wage distribution is described by (1). Expected cost expression (2) follows from integration and straightforward computations. ■

In this simple environment, the savings that can be obtained using random incentives are large: the cost of incentives goes from $\frac{\pi_A}{q}$ for deterministic mechanisms, to less that $\frac{\pi_A}{q}\frac{\pi_A}{k}$ for the optimal random incentive scheme. For instance, if the penalty for corruption is twice as high as the benefit of corruption, i.e. $k \geq 2\pi_A$, the principal would be able to save more than 50% on the cost of wages by using random incentives.[6]

This is of course a particularly simple environment. To properly assess the usefulness of random incentives we turn to a more general framework which allows for pre-existing asymmetric information, and more general bargaining structures.

3 Optimal incentives with pre-existing asymmetric information

3.1 Framework

Our more general framework coincides with that of Section 2 but extends it in three important ways:

- the agent's benefit π_A from corruption is now private information to the agent, distributed according to c.d.f. F_{π_A};
- the monitor now has a privately observed cost $\eta \geq 0$ for accepting a bribe, distributed according to c.d.f. F_η with density f_η;

[6] Note that gains remain large even if we consider simpler schemes: for the optimal *binary* wage distribution, the share of costs saved using random incentives will be exactly equal to $1 - \pi_A/k$. Indeed, the optimal binary wage distribution puts probability $1 - \pi_A/k$ on $w = 0$ and probability π_A/k on $w = \pi_A/q$.

- at the collusion stage, bargaining takes the form of probabilistic take-it-or-leave-it offers; the agent is the proposer with probability λ while the monitor proposes with probability $1-\lambda$.[7]

Altogether, payoffs now take the form

$$\begin{aligned}
u_P &= \pi_P \times c \quad -\gamma_w \times w - \gamma_q \times q \\
u_A &= \pi_A \times c \qquad\qquad\qquad\qquad -k \times c \times m \qquad\quad -\tau \\
u_M &= w \qquad\qquad\qquad\qquad\qquad -[q \times w + \eta] \times \mathbf{1}_{m \neq c} \quad +\tau.
\end{aligned}$$

The only difference from payoffs given in Section 2 is that the monitor now experiences an expected loss $qw + \eta$ rather than just qw when accepting a bribe, where η is a positive private cost of accepting bribes. There is asymmetric information over π_A, and η, but we maintain the assumption that parameters k, λ and q are known to the principal. We relax this assumption further in Section 4.

It is useful to note that the following extension of Fact 1 holds.

Fact 3. *If no agreement is reached at the collusion stage, the monitor's optimal continuation strategy is to send truthful reports $m = c$.*[8]

If the monitor acts as a proposer at the collusion stage, she demands a bribe $\tau \geq k$ when the agent is corrupt, and a bribe $\tau = 0$ when the agent is non-corrupt.

The agent accepts any offer $\tau \leq k$ when she is corrupt and any offer $\tau = 0$ when she is not corrupt.

An immediate implication is that non-corrupt agents get a payoff equal to 0.

[7] See Appendix A for an extension to arbitrary bargaining mechanisms.

[8] Fact 3 relies on the assumption that the monitor cannot commit to sending false reports about a non-corrupt agent. We allow for such commitment power in Appendix A and show that it does not affect our main results.

Optimal policies under budget constraints. Given a distribution of wages F_w, a corrupt agent of type π_A gets an expected payoff

$$U_A(\pi_A) = \pi_A - k + \lambda \max_{\tau \in [0, \pi_A]} (k - \tau) \text{prob}(qw + \eta < \tau).$$

The agent will choose to be corrupt if and only if $U_A(\pi_A) > 0$. Note that $U_A(\pi_A)$ is increasing in π_A, so that given a wage profile, agents follow a threshold strategy. Given a distribution F_w, let us denote by $\overline{\pi}_A(F_w)$ the highest non-corrupt type.

The principal's optimization problem over wage distribution F_w can be decomposed as follows: first, given a budget w_0, find the distribution of wages F_w that maximizes threshold $\overline{\pi}_A(F_w)$ under budget constraint $\mathbb{E}_{F_w}[w] = w_0$ — this is the *corruption-minimizing* wage schedule, given budget w_0. The overall optimum can then be obtained by optimizing over budget w_0. We believe that this "fixed budget" version of the principal's problem is particularly amenable to practical implementation and reflects the constraints that real-life institutions frequently operate under. Field experimentation with random incentive schemes seems more likely to happen if it takes as given existing monitoring budgets.

3.2 When is additional asymmetric information desirable?

Definition 1. *We say that a wage profile with c.d.f. F_w is random if and only if the support of F_w contains at least two elements.*

Proposition 2 (ambiguous optimal policy)**.** *(i) Whenever F_η is strictly concave over the range $[0, k]$, the corruption-minimizing wage profile under any budget $w_0 > 0$ is random.*

(ii) Whenever F_η is strictly convex over the range $[0, k]$, the corruption-minimizing wage profile under any budget $w_0 > 0$ is deterministic.

To get some intuition for this result, consider the agent's payoff from taking action $c = 1$:

$$\begin{aligned} U_A(\pi_A) &= \pi_A - k + \lambda \max_{\tau \in [0, \pi_A]} (k - \tau) \text{prob}(qw + \eta < \tau) \\ &= \pi_A - k + \lambda \max_{\tau \in [0, \pi_A]} (k - \tau) \mathbb{E}_{F_w}[F_\eta(\tau - qw)]. \end{aligned}$$

If F_η is strictly convex over the support of $\tau - qw$, the agent's payoff from a random wage schedule is larger than her payoff from a deterministic one with the same expectation. If F_η is strictly concave over the support of $\tau - qw$, the agent's payoff from a random wage schedule is smaller than her payoff from a deterministic one with the same expectation.

If F_w is neither concave nor convex over $[0, k]$ we can still provide sufficient conditions for random wage profiles to be optimal. Fix a deterministic wage w_0 and denote by τ_0 the highest solution to a corrupt agent's optimal bribe problem when the monitor is compensated with a deterministic wage w_0,

$$\max_\tau (k - \tau) \text{prob}(qw_0 + \eta < \tau).$$

Proposition 3 (sufficient condition for random incentives)**.** *Whenever $\tau_0 \leq \frac{k}{2}$, the corruption-minimizing policy given budget w_0 is random.*

In words, if starting from a deterministic wage, the agent's optimal bribe is less than half the cost of prosecution, it is optimal to use random wages.

Because adding further asymmetric information does not necessarily improve incentive provision, correct policy design will necessary depend on the restrictions, subjective or objective, that the principal can impose on the environment. We believe that specifying beliefs is often difficult for principals, which makes practical design exercises difficult. We approach this problem in three ways. First we characterize the max-min wage distribution and show that it coincides with the benchmark wage distribution F_w^{bmk} that is optimal in the simple environment of Section 2. Second, we solve for the optimal policy when the principal has a Bayesian prior over the environment. Third, we show how to perform prior-free policy evaluations using unverified report data.

3.3 Max-min optimal policy design

Take as given a budget w_0. Recall that given a wage distribution F_w, a distribution of private costs F_η, and bargaining power λ, we denote by $\overline{\pi}_A(F_w)$ the highest value of benefit π_A such that the agent still chooses to be non-corrupt. This section treats environment F_η, λ as a choice variable for nature, and we emphasize that threshold $\overline{\pi}_A$ depends on F_η and λ by using the notation $\overline{\pi}_A(F_w, F_\eta, \lambda)$.

We ask what is the max-min corruption-minimizing wage distribution, i.e. the solution to

$$\max_{\substack{F_w \\ \text{s.t. } \mathbb{E}_{F_w}[w]=w_0}} \min_{F_\eta,\lambda} \overline{\pi}_A(F_w, F_\eta, \lambda).$$

Denote by $\overline{\pi}_A^0$ the highest non-corruption threshold affordable under budget w_0, when the cost of keeping an agent of type π_A non-corrupt is given by the benchmark cost function $W^{\mathsf{bmk}}(\cdot)$ defined in Proposition 1, i.e. let $\overline{\pi}_A^0$ be the unique solution to $W^{\mathsf{bmk}}(\overline{\pi}_A^0) = w_0$. The following result holds.

Proposition 4 (max-min optimal incentives)**.** *The max-min optimal level of non-corruption is*

$$\max_{\substack{F_w \\ s.t.\ \mathbb{E}_{F_w}[w]=w_0}} \min_{F_\eta,\lambda} \overline{\pi}_A(F_w, F_\eta, \lambda) = \overline{\pi}_A^0.$$

It is attained by using the benchmark wage distribution defined in Section 2: $F_w^{\mathsf{bmk}}(w) = \frac{k-\overline{\pi}_A^0}{k-qw}$. The worse case environment is also that of Section 2, i.e. it sets $F_\eta(0) = 1$ and $\lambda = 1$.

3.4 Bayesian optimal incentives

We now characterize optimal incentives in the case where F_η is concave over the range $[0, k]$. We know from Proposition 3 that the optimal policy uses random incentives. For simplicity we also assume that $[0, k]$ is included in the support of F_η.

To facilitate exposition, it is helpful to consider the dual problem of minimizing costs given a target corruption threshold π_A. For any budget w_0 one can then compute the highest threshold π_A whose incentive cost is affordable under w_0. Fix a target threshold π_A and a wage policy F_w. An agent of type π_A chooses to remain non-corrupt if and only if, for all possible bribes $\tau \in [0, \pi_A]$,

$$\begin{aligned} & \pi_A - k + \lambda(k-\tau)\text{prob}(\eta + qw < \tau) \leq 0 \qquad (4) \\ \iff & \text{prob}(\eta + qw < \tau) \leq \frac{k - \pi_A}{\lambda(k-\tau)}. \end{aligned}$$

Define

$$m_0 \equiv \min_{\tau \in [0,\pi_A]} \frac{k - \pi_A}{\lambda(k-\tau)\text{prob}(\eta < \tau)} \qquad (5)$$

and denote by τ_0 the highest solution to (5). Note that agents with type π_A such that $m_0 \geq 1$ choose to remain non-corrupt for any wage distribution.[9] We focus on agents of type π_A such that $m_0 < 1$.

Let $\overline{\tau} \equiv \frac{\pi_A - (1-\lambda)k}{\lambda}$ and note that $\overline{\tau} > \tau_0$ for all π_A such that $m_0 < 1$.[10] Denote by Φ the operator over c.d.f.s F such that for all $w \in [0, +\infty)$,

$$\Phi(F)(w) = \begin{cases} m_0 & \text{if } w \in [0, \frac{\tau_0}{q}], \\ \min\left\{1, \frac{k-\pi_A}{f_\eta(0)\lambda(k-qw)^2} - \int_0^{qw} \frac{f_\eta'(\widehat{\eta})}{f_\eta(0)} F\left(w - \frac{\widehat{\eta}}{q}\right) \mathrm{d}\widehat{\eta}\right\} & \text{if } w \in (\frac{\tau_0}{q}, \frac{\overline{\tau}}{q}), \\ 1 & \text{if } w \geq \frac{\overline{\tau}}{q}. \end{cases} \qquad (6)$$

[9] Indeed, $m_0 \geq 1$ implies $0 \geq \pi_A - k + \max_\tau \lambda(k-\tau)\text{prob}(\eta < \tau) \geq \pi_A - k + \max_\tau \lambda(k-\tau)\text{prob}(\eta + qw < \tau)$.

[10] Indeed, $m_0 < 1$ implies $\frac{k-\pi_A}{\lambda(k-\tau_0)} < 1 \iff \tau_0 < \overline{\tau}$.

Proposition 5 (Bayes-optimal incentives)**.** *Assume that F_η is concave over the range $[0, k]$. The optimal wage distribution F_w^* satisfies the following properties:*

(i) $\forall w \in [0, \tau_0/q]$*,* $F_w^*(w) = m_0$*;*

(ii) over the range $\tau \in [\tau_0, k]$*, incentive compatibility condition* (4) *holds with equality for all* τ *such that* $F_w^*(\tau/q) < 1$*;*

(iii) F_w^* *is the unique solution to fixed point equation* $F_w^* = \Phi(F_w^*)$; *furthermore,* Φ *is a contraction mapping under the sup norm.*

Point (*ii*) of Proposition 5 echoes Proposition 1. Incentive compatibility of non-corrupt behavior at every $\tau \in [0, \pi_A]$ implies a bound on the distribution of corruption costs $\eta + qw$. The intuition for point (*i*) comes from writing $\text{prob}(\eta + qw < \tau) = \text{prob}(\eta < \tau)F_w(0) + \text{prob}(\eta + qw < \tau | qw \in (0, \tau))\text{prob}(qw \in (0, \tau))$. This implies that m_0 is necessarily an upper bound to $F_w(0)$ and that whenever $F_w(0) = m_0$, F_w can place no mass on $(0, \tau_0/q)$.

4 Prior-free policy evaluation

Proposition 5 solved for the optimal wage profile for a class of well-behaved priors F_η under which the optimal policy is random. We now show that even if the principal is unwilling to specify a prior belief over the underlying environment, it is possible to perform prior-free local policy evaluations provided the principal has access to appropriate experimental data. Our inference results do not require the principal to know any of the parameters of the environment, in particular, the cost k imposed by the judiciary on corrupt agents, the likelihood q of detection, and bargaining power λ need not be known to the observer.

Given budget w_0, consider two policies F_w^0, F_w^1 such that $\mathbb{E}_{F_w^0}[w] = \mathbb{E}_{F_w^1}[w] = w_0$. For any $\epsilon \in [0, 1]$, denote by F_w^ϵ the mixture

$$F_w^\epsilon \equiv (1 - \epsilon)F_w^0 + \epsilon F_w^1.$$

Imagine that policy F_w^ϵ is implemented over an infinite population of exchangeable monitor and agent pairs. Denote by $m^\epsilon \in \{0,1\}$ equilibrium report from monitors, and by $c^\epsilon \in \{0,1\}$ the corruption decision of agents. For any statistic Z, we denote by $\widehat{\mathbb{E}}Z$ the population average of Z.

Given a policy F_w^ϵ, denote by $\overline{R}_\epsilon = \widehat{\mathbb{E}}[m^\epsilon]$ the proportion of monitors reporting corruption, and by $\overline{C}_\epsilon = \widehat{\mathbb{E}}[c^\epsilon]$ the proportion of agents that are corrupt. Our first result clarifies that starting from a deterministic wage, unconditional report data $\overline{R}_\epsilon$ is not a sufficient statistic to evaluate whether a policy change increases or reduces underlying corruption.

Fact 4 (unreliable aggregate reports)**.** *Consider a default deterministic wage w_0, and any alternative random incentive scheme F_w^1 such that $\mathbb{E}_{F_w^1}[w] = w_0$.*

Regardless of whether $\overline{R}_0 < \overline{R}_1$ or $\overline{R}_0 > \overline{R}_1$, there exist specifications of k, F_{π_A} and F_η such that $\overline{C}_0 > \overline{C}_1$, and specifications of k, F_{π_A} and F_η such that $\overline{C}_0 < \overline{C}_1$.

In words, the ordering of aggregate reports places no restrictions on the ordering of underlying corruption. Indeed, reports of corruption depend on both underlying rates of corruption, and the monitors' decisions to report corruption or not. Hence, a scheme that facilitates bribing the monitor may end up increasing actual corruption while decreasing aggregate reports of corruption.

Still, we now show that using *conditional* report data it is possible to evaluate local policy changes. A draw w from F_w^ϵ can always be decomposed as a draw from a Bernoulli variable $X \in \{0,1\}$ with prob$(X = 1) = \epsilon$, followed by a draw of w according to F_w^X. Define mean reports conditional on X by $R_\epsilon(X) \equiv \widehat{\mathbb{E}}\left[m^\epsilon | X\right]$.

Proposition 6 (prior-free policy evaluation)**.** *The impact of local policy changes on underlying corruption can be identified from observable conditional reports:*

$$\mathsf{sgn}\left[\frac{\partial \overline{C}_\epsilon}{\partial \epsilon}\right] = \mathsf{sgn}\left[R_\epsilon(X=0) - R_\epsilon(X=1)\right].^{11}$$

This implies that a small movement from F_w^0 to F_w^1 decreases corruption if and only if there are more reports of corruption conditional on $X = 1$ (i.e. when the wage is drawn according to F_w^1) than conditional of $X = 0$ (i.e. when the wage is drawn according to F_w^0).

An immediate corollary is that unverified report data from a single policy experiment lets us identify optimal local policy changes. Take as given a distribution of wages with density f_w^0 . Denote by $\mathcal{P}$ the set of alternative policies f_w^1 satisfying

$$\mathsf{supp}\, f_w^1 = \mathsf{supp}\, f_w^0 \quad \text{and} \quad \mathbb{E}_{f_w^0}[w] = \mathbb{E}_{f_w^1}[w].$$

For any such f_w^1, construct the mixture $f_w^\epsilon = (1-\epsilon) f_w^0 + \epsilon f_w^1$ and define

$$\nabla_{f_w^1} \overline{C} = \frac{\partial \mathbb{E}[c^\epsilon | f_w^\epsilon]}{\partial \epsilon}\Big|_{\epsilon=0}.$$

This measures the marginal change in corruption following a marginal move in the direction of f_w^1.

Since, f_w^0 and f_w^1 have the same support it is possible to construct a random variable X with values in $\{0, 1\}$ coupled with wage w so that $f_w^0(w|X = 1) = f_w^1(w)$. Indeed, simply draw X conditional on w according to a distribution of the form

$$\text{prob}(X = 1|w) = \lambda \frac{f_w^1(w)}{f_w^0(w)},$$

with λ small enough that $\text{prob}(X = 1|w) \leq 1$ for all w. Denote by $R_0(f_w^1) \equiv \mathbb{E}_{f_w^0}[m|X = 1]$ average reports of corruption under the synthetic distribution of wages $f_w^0(w|X = 1) = f_w^1(w)$.

[11] The result continues to hold even if only some share of agents can update its play to equilibrium following a change in policy.

Corollary 1 (optimal local policy change)**.** *The optimal local policy change in $\mathcal{P}$ at f_0 is determined from report data alone:*

$$\arg\min_{f_w^1 \in \mathcal{P}} \nabla_{f_w^1} \overline{C} = \arg\min_{f_w^1 \in \mathcal{P}} \overline{R}_0 - R_0(f_w^1).$$

This result is useful for several reasons. First it suggests a simple data-driven gradient descent algorithm to find corruption-minimizing policies. Second, with field experiment in minds, it suggests that one can propose a plausible alternative policy to a deterministic default by using report data from a partial equilibrium pilot that merely randomizes wages according to any full support distribution. This is reassuring since the space of policies is infinite dimensional, which makes trial and error policy search difficult.

5 Discussion

5.1 Summary

We study incentive provision in a principal-agent-monitor model in which the agent and monitor can collude on what message to send to the principal. We explore the idea that since collusion is a side-contracting problem, it may be addressed by introducing asymmetric information frictions that make contracting difficult. Indeed, by using random incentives that serve as the monitor's private information, the principal can decrease the rents that the agent extracts from contracting with the monitor. This can result in significant cost reductions over deterministic incentive schemes. In a benchmark environment with no pre-existing information, random incentives reduce the cost of incentive provision by over 50% for plausible parameter values.

In the presence of pre-existing asymmetric information taking the form of idiosyncratic costs for accepting bribes, introducing additional asymmetric information may or may not be optimal. We provide sufficient conditions for random incentives to be optimal or not as a function of the concavity or convexity of the c.d.f. of idiosyncratic bribery costs. Furthermore, we are able to show that although aggregate reports of corruption do not provide a reliable indicator of underlying corruption, it is possible to evaluate local policy changes on the basis of unverified report data alone. This provides an experiment-ready framework to test the effectiveness of random incentives in reducing corruption and collusion in organizations.

5.2 Extensions

Our framework obviously admits many plausible extensions. We briefly describe a few and delineate the way our results extend in each case. Formal treatment of these extensions is delayed to Appendix A.

Multiple monitors. Section 2 shows that the effectiveness of deterministic incentive schemes may be undermined by the possibility of collusion. In this case endogenous asymmetric information may significantly reduce the cost of incentive provision. This point is robust to the introduction of multiple monitors. Indeed, while cross-checking the messages of different monitors using mechanisms à la Maskin (1999) successfully reveals public information in the absence of collusion (see Duflo et al. (2013) for a recent field implementation), such mechanisms are fragile to the possibility of collusion: monitors can collude on what message to send. In the example without pre-existing private information described in Appendix A the cost of bribing two monitors turns out to be no higher than the cost of bribing a single monitor with twice the incentives. As a result, asymmetric information also emerges as an effective strategy to reduce monitoring costs.[12]

Extortion. In the model of Sections 2 and 3, we assume that the monitor sends a subgame perfect message following disagreement at the side-contracting stage. This implies that the

[12] Note that in the presence of endogenous or exogenous asymmetric information, there may indeed be efficiency gains from using multiple monitors.

monitor can never extract bribes from an agent which she observes to be non-corrupt. As Olken and Pande (2012) highlight, this prediction is frequently invalidated: honest agents are often extorted bribes by monitors. A variation of our baseline model naturally accounts for this. Assume that when she has the bargaining power, the monitor is able to commit to a message she would send in the event of a bargaining failure. Assume also and that even non-corrupt monitors experience a cost when they are reported as corrupt. A monitor can then extract rents from honest agents by committing to report the agent as corrupt unless a bribe is paid. While this changes the agent's incentives to be corrupt, we show in Appendix A that our main results continue to hold in this setting: random incentives may reduce the cost of incentive provision, and it is possible to perform local policy evaluation on the basis of conditional report data only.

Dynamic incentives. The model of Sections 2 and 3 is static. Realistically however, wages w may represent the present discounted value of future wages which the monitor stands to lose, should she be fired. One potential difficulty with dynamic extensions to our framework is that the continuation value of the monitor would depend on her ability to raise bribes from agents, so that incentives for truth-telling would in fact depend on the rents obtained from bribes. While it is reasonable to expect that our basic qualitative message would survive in some form, it less obvious that our stronger results, and especially the prior-free policy evaluation property described in Proposition 6 would extend. Remarkably, we are able to show in Appendix A that whenever the monitor's type η is persistent, Proposition 6 extends as is.

5.3 Implementation

Because the cost-savings from random incentives are significant in plausible environments, and because the fragility of counter-corruption schemes to collusion is increasingly recognized as a first-order issue, we believe that the policy recommendation that emerges from our analysis is an attractive candidate for field implementation. We describe below how we envision running such an exercise.

Wages versus scrutiny. Randomizing wages has distributional implications which stakeholders may find unfair. However, as we noted in Section 2 the monitor's incentives for truth-telling are captured by her *expected* lost wages qw from misreporting. Although we chose to focus on wages w as a policy instrument, our analysis would be unchanged if the intensity of scrutiny q was the policy instrument of interest. Since changing q does not affect the welfare of the monitor when she reports truthfully, it does not have the adverse distributional consequences of random wages in equilibrium. For this reason, varying the level of scrutiny imposed on monitors may be a more suitable policy instrument for practical implementation. For instance, in public infrastructure projects where, as in Olken (2007), local officials play the role of natural monitors, one may vary the probability with which the project gets audited by an external, well compensated engineering firm.

Picking a candidate policy. Proposition 6 provides a framework for policy evaluation using report data. One difficulty in setting up a field implementation of random incentive schemes is to construct a plausible policy alternative to deterministic incentives. Distributions of wages are high dimensional objects and absent great luck, simple trial and error seems unlikely to succeed. Fortunately, as we highlighted in Section 4, Corollary 1 provides guidance on what alternative policy to choose using report data from any random incentive trial, provided it has a sufficiently rich support: choose the distribution that maximizes reports of corruption keeping average incentives constant. This implies that one can form a plausible candidate policy using report data from a pilot intervention using any arbitrary full support distribution of wages.

Continuous evaluation. Proposition 6 provides a framework for local policy evaluation. Interestingly, it can be used for global policy evaluation provided the policy is phased in progressively, i.e. by progressively increasing the proportion of monitors placed under the new incentive scheme, and recording report data as the policy is being implemented. Once reports suggest that there are no longer any local improvements, phasing-in of the alternative policy may be stopped.

Appendix

A Extensions

A.1 Collusion with Multiple Monitors

This extension illustrates how collusion can undermine the effectiveness of deterministic incentive schemes even when the principal can use multiple monitors to cross-check their reports. We consider a principal who hires two monitors, $i = 1, 2$, to check the agent. As in the model of Section 2, the agent takes a corruption decision $c \in \{0, 1\}$, where $c = 1$ gives the agent a benefit π_A and comes at a cost $\pi_P < 0$ to the principal. The agent's action is not observable to the principal, but is observed by both monitors. After observing the agent's action, each monitor $i = 1, 2$ sends a report $m_i \in \{0, 1\}$ to the principal. Report $m_i = 1$ by either monitor triggers an exogenous judiciary process that imposes an expected cost $k > \pi_A$ on corrupt agents and (for simplicity) a cost of 0 on non-corrupt agents.

The principal detects false reports $m_i \neq c$ with probability $q \in (0, 1)$. If both monitors send the same report and the principal does not find evidence of misreporting, then both monitors are paid their wage w. If both monitors send different reports and the principal does not find evidence of misreporting, the monitor reporting $m = 0$ gets fired and the other monitor gets her wage w. If the principal finds evidence that a report was false, the monitor sending that report gets fired.

The timing of the game is as follows:

1. the principal offers a fixed wage w to each monitor;

2. the agent makes a corruption decision $c \in \{0, 1\}$;

3. under *collusion*, the agent sequentially makes take-it-or-leave-it bribe offers τ_1 and τ_2 to monitors 1 and 2 in exchange for sending message $m_i = 0$, which each monitor accepts or rejects — we assume perfect commitment so that whenever a monitor accepts the bribe, she does send message $m = 0$; under *no-collusion* nothing occurs;

4. under *no-collusion* or, under *collusion* if there was no agreement between the agent and monitor i in the previous stage, monitor i sends message m_i maximizing her final payoff.

The following result generalizes Fact 2 to the current setting.

Fact A.1. *Assume that the principal hires two monitors and uses deterministic wages. Under* no collusion *the principal can induce the agent to be non-corrupt at 0 cost.*

Under collusion*, the minimum cost of wages needed to induce the agent to be non-corrupt is equal to* $\frac{\pi_A}{q}$.

Proof. Under *no collusion*, it is an equilibrium for both monitors to send a truthful report for any wage $w > 0$. Under this equilibrium, the payoff that the agent gets when corrupt is $\pi_A - k < 0$, while her payoff when non-corrupt is 0.[13]

Consider next the case of *collusion*. Solving the game by backward induction, if a corrupt agent successfully bribed the first monitor, then monitor 2 accepts a bribe τ_2 if and only $\tau_2 > qw$. If the first monitor expects that the agent will successfully bribe the second monitor, she accepts a bribe τ_1 if and only if $\tau_1 > qw$. The payoff of a corrupt agent who bribes both monitors is $\pi_A - 2qw$. The payoff of a non-corrupt agent is 0, so the agent will be non-corrupt if and only if $\pi_A - 2qw \leq 0$, or $w \geq \frac{\pi_A}{2q}$. Therefore, the minimum cost of wages needed to induce the agent to be non-corrupt is $\frac{\pi_A}{q}$. ■

[13] Note that, when $q < \frac{1}{2}$, there is also an equilibrium in which both monitors send message $m = 1$ regardless of the agent's behavior or their wage.

A.2 Arbitrary bargaining

The model in the main text simplifies the side-contracting stage by assuming take-it-or-leave-it offers. This appendix extends the model by considering more general bargaining structures. We study a model that in which the monitor and the agent can use any individually rational and incentive compatible mechanism at the side-contracting stage, but that is otherwise identical to the basic model in Section 2.

By the revelation principle, we can restrict attention to mechanisms under which the monitor announces her private information (i.e., her wage) and this announcement determines the bargaining outcome. Such a bargaining mechanism is characterized by two functions: (i) $P(w)$, the probability with which monitor and agent reach an agreement when the monitor's wage is w; and (ii) $\tau(w)$, the expected 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. If there is no agreement, the monitor sends the message that maximizes her final payoff (i.e., she sends a truthful message).

Given a wage schedule F and a mechanism (P, τ), the agent's expected payoff from being corrupt is $U_A = \pi_A - k + \int (P(w)k - \tau(w))\, dF(w)$. The individual rationality constraint of a corrupt agent is $U_A \geq \pi_A - k$, since a corrupt agent can guarantee $\pi_A - k$ by not participating in the mechanism.

The payoff that a monitor with wage w who announces wage w' gets under mechanism (P, τ) when the agent is corrupt is $\tilde{U}_M(w, w') = \tau(w') + (1 - P(w')q)w$. By incentive compatibility, $U_M(w) \equiv \tilde{U}_M(w, w) \geq \tilde{U}_M(w, w')$ for all $w' \neq w$. By individual rationality, $U_M(w) \geq w$ for all w, since a monitor with wage w obtains a payoff of w by not participating in the mechanism and sending a truthful report.

Given a mechanism (P, τ) and a wage distribution F, the weighted sum of the agent's and monitor's payoff when the monitor is corrupt is

$$(1-\lambda)\int U_M(w)dF(w)+\lambda U_A, \tag{7}$$

where the weight $\lambda \in [0,1]$ represents the monitor's bargaining power. For every wage schedule F and every $\lambda \in [0,1]$, let $\Gamma(F,\lambda)$ be the set of incentive compatible and individually rational bargaining mechanisms that maximize (7). We assume that, at the side-contracting stage, the monitor and the agent use a bargaining mechanism in $\Gamma(F,\lambda)$. Let $\tilde{U}_A(F,\lambda)$ be the lowest utility that a corrupt agent gets under a bargaining mechanism in $\Gamma(F,\lambda)$. The agent has an incentive to be non-corrupt if $\tilde{U}_A(F,\lambda) \leq 0$.

The following result generalizes Proposition 1 to this setting.

Proposition A.1. *Suppose that, at the collusion stage, the monitor and the agent use an incentive compatible and individually rational mechanism that maximizes* (7)*.*

(i) If $\lambda \in (1/2,1]$, the cost minimizing wage distribution F_w^{gen} that induces the agent not to be corrupt is described by

$$\forall w \in [0,\pi_A/q], \quad F_w^{gen}(w)=\left(\frac{k-\pi_A}{k-qw}\right)^{\frac{2\lambda-1}{\lambda}}. \tag{8}$$

(ii) If $\lambda \in [0,1/2]$, the cost minimizing wage distribution F_w^{gen} that induces the agent not to be corrupt has $F_w^{gen}(0)=1$.

Proof. By standard arguments, any incentive compatible mechanism (P,τ) must satisfy: (i) $P(w)$ is decreasing, and (ii) $U_M'(w)=1-qP(w)$. This last condition and the monitor's individual rationality constraint (i.e., $U_M(w)\geq w$ for all w) imply that $U_M(w)=\int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}+w+c$ for some constant $c\geq 0$ (where $\overline{w}$ is the highest wage in the support of F). Since $U_M(w)=\tau(w)+(1-qP(w))w$, $\tau(w)=P(w)qw+\int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}+c$. The weighted sum of payoffs when the agent is corrupt is

$$\begin{aligned}&(1-\lambda)\int_{\underline{w}}^{\overline{w}} U_M(w)dF(w)+\lambda U_A\\ &=\int_{\underline{w}}^{\overline{w}}\left[(1-\lambda)(\tau(w)+(1-qP(w))w)+\lambda(P(w)k-\tau(w))\right]dF(w)+\lambda(\pi_A-k)\\ &=\int_{\underline{w}}^{\overline{w}}\left[P(w)\lambda\left(k-qw\right)+(1-\lambda)w\right]dF(w)+\lambda(\pi_A-k)+(1-2\lambda)\left(\int_{\underline{w}}^{\overline{w}} qP(w)F(w)dw+c\right).\end{aligned} \tag{9}$$

We use the following lemma.

Lemma A.1. *For all $\lambda\in(1/2,1]$, the mechanism (P,τ) that maximizes (9) has: (i) $P(w)=1$ if $w<w^*$ and $P(w)=0$ if $w>w^*$ for some w^*, and (ii) $\tau(w)=P(w)qw+\int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}$.*

Proof. We first show that the mechanism that maximize (9) is such that $P(w)$ only takes values 0 or 1. Suppose by contradiction that there exists an interval V such that $P(w)\in(0,1)$ for all $w\in V$, and let $H\equiv\int_V\lambda(k-qw)dF(w)+(1-2\lambda)\int_V qF(w)dw$. If $H\geq 0$, increasing $P(w)$ over this interval (subject to the constraint that P is decreasing) makes (9) larger. If $H<0$, decreasing $P(w)$ over this interval (subject to the constraint that P is decreasing) also makes (9) larger. Such improvements are exhausted when $P(w)$ only takes values 0 and 1. Since $P(\cdot)$ is decreasing, when $P(\cdot)$ only takes values 0 or 1 there must exist a wage w^* such that $P(w)=1$ if $w<w^*$ and $P(w)=0$ if $w>w^*$. Finally, (9) is maximized by setting $c=0$ when $\lambda\in(1/2,1]$, so $\tau(w)=P(w)qw+\int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}$. ■

We now conclude the proof of Proposition A.1, begining with point (i). Fix $\lambda\in(1/2,1]$ and let (P,τ) be the mechanism that maximizes (9). By Lemma A.1, $P(w)=1$ if $w<w^*$ and $P(w)=0$ if $w>w^*$ for some w^* and $c=0$. Under this mechanism (9) becomes

$$\lambda\left[F(w^*)k-\int_0^{w^*} qwdF(w)+\pi_A-k\right]+(1-\lambda)\int wdF(w)+(1-2\lambda)\int_0^{w^*} qF(w)dw.$$

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

$$\begin{aligned} & \lambda\left[F(w^*)k - \int_0^{w^*} qwdF(w)\right] + (1-2\lambda)\int_0^{w^*} qF(w)dw \\ \geq\ & \lambda\left[F(\hat{w})k - \int_0^{\hat{w}} qwdF(w)\right] + (1-2\lambda)\int_0^{\hat{w}} qF(w)dw. \end{aligned} \tag{10}$$

Otherwise, if (10) did not hold for some $\hat{w} \neq w^*$, the weighted sum of payoffs would be strictly larger under mechanism $(\hat{P}, \hat{\tau})$ with $\hat{P}(w) = 1$ if $w < \hat{w}$ and $\hat{P}(w) = 0$ if $w > \hat{w}$.

Consider next the principal's problem, who chooses a wage schedule F to minimize expected wage payments subject to the constraint that the agent has an incentive to be non-corrupt. By first order stochastic dominance, it is cheaper for the principal to choose a wage schedule F such that (10) holds with equality for all $\hat{w}$ such that $F(\hat{w}) < 1$; that is, under the optimal wage schedule F the right-hand side of (10) is constant for all $\hat{w}$ such that $F(\hat{w}) < 1$. Differentiating the right-hand side of (10) with respect to $\hat{w}$,

$$F'(\hat{w})\lambda[k - q\hat{w}] + qF(\hat{w})(1-2\lambda) = 0. \tag{11}$$

The solution to the differential equation (11) is $F(w) = C\left(\frac{1}{k-qw}\right)^{\frac{2\lambda-1}{\lambda}}$ for some constant C.

We now determine the value of the constant C. For any $\hat{w}$ in the support of F, let $(P_{\hat{w}}, \tau_{\hat{w}})$ be the mechanism with $P_{\hat{w}}(w) = \mathbf{1}_{\{w \leq \hat{w}\}}$ and $\tau_{\hat{w}}(w) = P_{\hat{w}}(w)qw + \int_w^{\overline{w}} qP_{\hat{w}}(\tilde{w})d\tilde{w} = \mathbf{1}_{\{w \leq \hat{w}\}}q\hat{w}$. Since (10) holds with equality for all $\hat{w}$ under the optimal distribution, all such mechanisms maximize (9); that is, all such mechanisms are in $\Gamma(F,\lambda)$. Recall that $\tilde{U}_A(F,\lambda)$ is the lowest utility that the agent gets under a mechanism in $\Gamma(F,\lambda)$, and that the agent has an incentive to be non-corrupt only if $\tilde{U}_A(F,\lambda) \leq 0$. The agent's utility under mechanism $(P_{\hat{w}}, \tau_{\hat{w}})$ is $u(\hat{w}) = F(\hat{w})(k - q\hat{w}) + \pi_A - k$. Note that $u'(\hat{w}) = F'(\hat{w})(k - q\hat{w}) - qF(\hat{w}) = qF(\hat{w})[\frac{2\lambda-1}{\lambda} - 1] \leq 0$, where the second equality follows since F satisfies (11). Therefore, the lowest utility that the agent gets under a mechanism in $\Gamma(F,\lambda)$ is $\tilde{U}_A(F,\lambda) = u(\overline{w}) = k - q\overline{w} + \pi_A - k = \pi_A - q\overline{w}$, where $\overline{w}$ is the highest wage in the support of F; i.e., $\overline{w}$ is such that $F(\overline{w}) = C\left(\frac{1}{k-q\overline{w}}\right)^{\frac{2\lambda-1}{\lambda}} = 1$. The agent does not have an incentive to corrupt if $\tilde{U}_A(F,\lambda) = u(\overline{w}) \leq 0$, or $\overline{w} \geq \frac{\pi_A}{q}$. To minimize expected wages it is optimal to set $\overline{w} = \frac{\pi_A}{q}$. This implies $C = (k - \pi_A)^{\frac{2\lambda-1}{\lambda}}$, so the optimal distribution is (8).

We now turn to point (ii). When $\lambda \leq 1/2$, the mechanism (P,τ) that maximizes (9) must make the constant c as large as possible, subject the agent's IR constraint; that is, subject to $\pi_A - k + \int[P(w)k - \tau(w)]dF(w) \geq \pi_A - k$. Recall that $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} + c$. The maximum is achieved by choosing c such that $\int[P(w)k - \tau(w)]dF(w) = 0$. Therefore, for $\lambda \leq 1/2$ the agent's payoff from being corrupt under a mechanism that maximizes (9) is $\pi_A - k < 0$, regardless of the wage schedule. This implies that the agent has an incentive to be non-corrupt even when F has all its mass at $w = 0$. ■

A.3 Extortion

This section shows how our results extend to settings in which the monitor can extort a non-corrupt agent by committing to send a false report. The framework we consider is essentially the same as in Section 3. The only difference is that a monitor who makes an offer at the side-contracting stage can commit to sending a false report if the agent rejects her proposal. A report $m = 1$ triggers an exogenous judiciary process that imposes an expected cost $k > \pi_A$ on corrupt agents and an expected cost $k_0 \in (0, k]$ on non-corrupt agents.

Fact A.2. *If the monitor acts as proposer when the agent is non-corrupt, she demands a bribe $\tau = k_0$ if her type is $\eta < k_0$, and she demands no bribe (i.e., she demands $\tau = 0$) if her type is $\eta \geq k_0$. A non-corrupt agent accepts any offer $\tau \leq k_0$.*

Proof. Suppose the monitor makes an offer τ to a non-corrupt agent and commits to sending a false message if her proposal is rejected. In this case, it is optimal for a non-corrupt agent to accept the offer if and only if $\tau \leq k_0$: her payoff from accepting such an offer is $-\tau$, while her payoff from rejecting the offer is $-k_0$. The monitor's payoff from making an offer

$\tau \in (0, k_0]$ is $\tau - \eta$, while her payoff from not demanding a bribe is 0. A type η monitor finds it optimal to make an offer $\tau = k_0$ if only if $\eta < k_0$. ■

Fact A.3. *If the monitor acts as a proposer at the collusion stage, she demands a bribe $\tau \geq k$ when the agent is corrupt. A corrupt agent accepts any offer $\tau \leq k$.*

Proof. The proof of Fact A.3 is identical to the proof of Fact 3. ■

Fact A.2 implies that the payoff of a non-corrupt agent is $-(1-\lambda)k_0F_\eta(k_0)$, while Fact A.3 implies that the payoff of a corrupt agent of type π_A is $\pi_A - k + \lambda \max_\tau (k-\tau)\text{prob}(qw+\eta < \tau)$. Therefore, when the monitor can commit to sending a false report, an agent of type π_A will take action $c = 0$ if only if

$$\pi_A - (k - (1-\lambda)k_0F_\eta(k_0)) + \lambda \max_{\tau \in [0,k]} (k-\tau)\text{prob}(qw + \eta < \tau) \leq 0.$$

From the principal's perspective, the possibility of extortion by the monitor reduces the effective punishment cost that a corrupt agent incurs when the monitor sends report $m = 1$ down to $k - (1-\lambda)k_0F_\eta(k_0)$. With this modification all the results in Sections 3 and 4 continue to hold when the monitor can commit to sending a false message.

A.4 Dynamic incentives

The model in the main text assumes that the principal provides incentives to monitors by paying them a wage of zero if there is evidence that the monitor misreported. This appendix extends our analysis to settings in which the principal hires the monitor for multiple periods and in which a monitor who is found misreporting is fired and losses her continuation value of employment. The goal of this section is to show that, in this setting, we can still identify the impact of local policy changes using data from unverified reports.

Consider a principal who needs to repeatedly audit a population of agents. The principal hires a population of monitors to check the agents at each of infinitely many periods. Monitors are randomly matched with agents at each period. At time $t = 0$ the principal commits to a distribution of wages F_w and draws a wage w for each monitor from this distribution. This wage is observed by the monitor and not by the agents. Each monitor's wage is persistent: the monitor receives a constant wage at every period at which she is employed. Monitors have a persistent cost η from accepting a bribe, where η is distributed according to F_η. Within each period the structure of the game is the same as that in Section 3; the only difference is that a monitor who is found misreporting receives her current period wage w but losses her continuation value from employment.

Let $W(w, \eta)$ be the value function from maintaining employment of a monitor with wage w and type η, and normalize the monitor's value of unemployment to zero. The net benefit that a monitor with wage w and type η gets from accepting bribe τ from a corrupt agent is $(1 - \delta)(\tau - \eta) - q\delta W(w, \eta)$, where $\delta < 1$ is the discount factor. This implies the following.

Fact A.4. *A monitor with wage w and type η accepts an offer τ at the collusion stage if and only if $\tau > \eta + q\frac{\delta}{1-\delta}W(w, \eta)$.*

The next observation is the counterpart of Fact 3 to the current setting.

Fact A.5. *If no agreement is reached at the collusion stage, the monitor's optimal continuation strategy is to send truthful reports $m = c$.*

If the monitor acts as a proposer at the collusion stage, she demands a bribe $\tau \geq k$ when the agent is corrupt, and a bribe $\tau = 0$ when the agent is non-corrupt.

The agent accepts any offer $\tau \leq k$ when she is corrupt and any offer $\tau = 0$ when she is not corrupt.

The payoff of a non-corrupt agent is 0. Moreover, by Facts A.4 and A.5 the payoff of a corrupt agent of type π_A is

$$U_{\pi_A} = \pi_A - k + \lambda \max_{\tau \in [0,k]} (k - \tau) \text{prob} \left(q \frac{\delta}{1-\delta} W(w, \eta) + \eta < \tau \right).$$

We allow different agents to derive a different benefit π_A from being corrupt, and assume that the distribution of corruption benefits π_A is constant across periods. As in Section 4, let $\overline{C} = \widehat{\mathbb{E}}[c]$ be the proportion of agents who are corrupt.

Lemma A.2. *Let $\tau^* \leq k$ be the offer that corrupt agents make. A monitor with type η and wage w accepts offer τ^* from a corrupt agent if and only if $\eta < \overline{\eta}(\tau^*, w, \overline{C})$, where*

$$\overline{\eta}(\tau, w, \overline{C}) \equiv \frac{\tau(1 - \delta + \delta q (1-\lambda)\overline{C}) - q\delta w - q\delta(1-\lambda)k\overline{C}}{1-\delta}.$$

Proof. Consider a monitor with wage w and type η who is indifferent between accepting and rejecting an offer $\tau^* \leq k$ by a corrupt agent. The value function of this monitor is

$$\begin{aligned} W(w,\eta) &= (1-\delta)w + \delta W(w,\eta) + (1-\lambda)\overline{C}\left((1-\delta)(k-\eta) - q\delta W(w,\eta)\right) \Rightarrow \\ W(w,\eta) &= \frac{(1-\delta)(w + (1-\lambda)(k-\eta)\overline{C})}{1-\delta+\delta q(1-\lambda)\overline{C}}. \end{aligned} \tag{12}$$

The last term in the first expression is the payoff that the monitor gets when she is proposer against a corrupt agent.[14] Since this monitor is indifferent between accepting offer τ^* or rejecting it, $(1-\delta)(\tau^* - \eta) = q\delta W(w,\eta)$, which by equation (12) implies $\eta = \overline{\eta}(\tau^*, w, \overline{C})$. Monitors with wage w and type η such that $\eta < \overline{\eta}(\tau^*, w, \overline{C})$ find it optimal to accept τ^*, and monitors with wage w and type η such that $\eta > \overline{\eta}(\tau^*, w, \overline{C})$ find it optimal to reject τ^*. ■

We now show how Proposition 6 extends to this setting. Fix a budget w_0 and consider two policies F_w^0, F_w^1 such that $\mathbb{E}_{F_w^0}[w] = \mathbb{E}_{F_w^1}[w] = w_0$. Define F_w^ϵ as the mixture $F_w^\epsilon \equiv$

[14] Since this monitor is indifferent between accepting or rejecting offer $\tau^* \leq k$, she finds it (at least weakly) optimal to ask for a bribe k when making offers against a corrupt agent.

$(1-\epsilon)F_w^0 + \epsilon F_w^1$. As in Section 4, we imagine a draw w from F_w^ϵ as first drawing a Bernoulli variable $X \in \{0,1\}$ with prob$(X=1)=\epsilon$, followed by drawing w according to F_w^X. Denote by $m^\epsilon \in \{0,1\}$ the equilibrium report from monitors and by $c^\epsilon \in \{0,1\}$ the corruption decision of agents under policy F_w^ϵ. Recall that $R_\epsilon(X) = \widehat{\mathbb{E}}\left[m^\epsilon | X\right]$ are the mean reports conditional on X and that $\overline{C}_\epsilon = \widehat{\mathbb{E}}\left[c^\epsilon\right]$ is the proportion of agents who are corrupt under policy F_w^ϵ.

Proposition A.2. *The impact of local policy changes on underlying corruption can be identified from observable conditional reports:*

$$\mathsf{sgn}\left[\frac{\partial \overline{C}_\epsilon}{\partial \epsilon}\right] = \mathsf{sgn}\left[R_\epsilon(X=0) - R_\epsilon(X=1)\right].$$

Proof. Let τ_ϵ be the optimal offer by a corrupt agent under policy F_w^ϵ. By Lemma A.2, the probability that a monitor accepts this offer is prob $\left(\frac{q\delta W(w,\eta)}{1-\delta} + \eta < \tau_\epsilon\right)$ = prob $\left(\eta < \overline{\eta}(\tau_\epsilon, w, \overline{C})\right) = \mathbb{E}_{F_w}[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}))]$. The payoff of a corrupt agent with type π_A under policy F_w^ϵ is

$$U_{\pi_A}^\epsilon \quad = \quad \pi_A - k + \lambda(k-\tau_\epsilon)\left[(1-\epsilon)\mathbb{E}_{F_w^0}\left[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon))\right] + \epsilon \mathbb{E}_{F_w^1}\left[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon))\right]\right].$$

By the Envelope Theorem,

$$\begin{aligned}\frac{\partial U_{\pi_A}^\epsilon}{\partial \epsilon} \quad &= \quad (k-\tau_\epsilon)\left[\mathbb{E}_{F_w^1}\left[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon))\right] - \mathbb{E}_{F_w^0}\left[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon))\right]\right] \\ &\quad\quad + (k-\tau_\epsilon)\mathbb{E}_{F_w^\epsilon}\left[f_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon)) \times \frac{\partial \overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon)}{\partial \overline{C}_\epsilon}\frac{\partial \overline{C}_\epsilon}{\partial \epsilon}\right].\end{aligned}$$

The equation above can be written as

$$\begin{aligned}&\frac{\partial U_{\pi_A}^\epsilon}{\partial \epsilon} - (k-\tau_\epsilon)\mathbb{E}_{F_w^\epsilon}\left[f_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon)) \times \frac{\partial \overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon)}{\partial \overline{C}_\epsilon}\frac{\partial \overline{C}_\epsilon}{\partial \epsilon}\right] \\ &\quad\quad = (k-\tau_\epsilon)\left[\mathbb{E}_{F_w^1}\left[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon))\right] - \mathbb{E}_{F_w^0}\left[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon))\right]\right] \qquad (13)\end{aligned}$$

Note that $\frac{\partial \overline{\eta}(\tau_\epsilon, w, \overline{C}_\epsilon)}{\partial \overline{C}_\epsilon} = \frac{(\tau_\epsilon - k)\delta q(1-\lambda)}{1-\delta} \leq 0$. Since $\mathsf{sgn}\left[\frac{\partial U^\epsilon_{\pi_A}}{\partial \epsilon}\right] = \mathsf{sgn}\left[\frac{\partial \overline{C}^\epsilon}{\partial \epsilon}\right]$, the sign of the expression in the left-hand side of (13) is equal to $\mathsf{sgn}\left[\frac{\partial \overline{C}^\epsilon}{\partial \epsilon}\right]$. The proposition follows since $\mathsf{sgn}\left[\mathbb{E}_{F^1_w}\left[F_\eta(\overline{\eta}(\tau, w, \overline{C}_\epsilon))\right] - \mathbb{E}_{F^0_w}\left[F_\eta(\overline{\eta}(\tau, w, \overline{C}_\epsilon))\right]\right] = \mathsf{sgn}\left[R_\epsilon(X=0) - R_\epsilon(X=1)\right]$. ■

B Proofs

B.1 Proofs for Section 2

Proof of Fact 1. Under *collusion*, the monitor's payoff from accepting an offer τ from a corrupt agent is $\tau + (1-q)w$. Her payoff from rejecting the offer from a corrupt agent and sending message $m = 1$ is w. The monitor accepts the offer if and only if $\tau > qw$.

Under *no-collusion*, or if the monitor rejects the agent's offer, the monitor's payoff from sending message $m = c$ is w. Her payoff from sending a false message $m \neq c$ is $(1-q)w$, so the monitor has an incentive to send a truthful report for any wage $w \geq 0$.

Note that the expected payoff that a corrupt agent gets under collusion is $\pi_A - k + \max_\tau (k-\tau)\text{prob}(qw < \tau)$, while her payoff from not being corrupt is 0. If the agent expects to make a bribe offer $\tau > \pi_A$, her payoff from being corrupt is $\pi_A - k + (k-\tau)\text{prob}(qw < \tau) < 0$, so she would strictly prefer to be non-corrupt. ■

Proof of Fact 2. By Fact 1, given any wage w, under *no-collusion* the monitor's optimal strategy is to send a truthful report. The agent's payoff from action $c = 1$ is then $\pi_A - k < 0$ and her payoff from action $c = 0$ is 0. Thus, under *no-collusion* the principal can induce the agent to be non-corrupt at zero cost.

Consider next a setting with *collusion*. By Fact 1, the monitor accepts a bribe τ from a corrupt agent if and only if $\tau > qw$. The agent's payoff from taking $c = 1$ is therefore $\pi_A - \min\{k, qw\}$, while her payoff from action $c = 0$ is 0. It follows that the principal can induce the agent to take action $c = 0$ by setting a deterministic wage $w = \frac{\pi_A}{q}$. ■

B.2 Proofs for Section 3

Proof of Fact 3. If there is no agreement at the collusion stage the monitor's payoff from sending message $m = c$ is w. Her payoff from sending message $m \neq c$ is $(1-q)w$, so the monitor has an incentive to send a truthful report.

Consider next a monitor who acts as proposer at the collusion stage when the agent is corrupt. Note that a corrupt agent accepts any offer $\tau \leq k$: her payoff from accepting such an offer is $\pi_A - \tau$, while her payoff from rejecting the offer is $\pi_A - k$. The monitor's payoff from making an offer $\tau \leq k$ is then $\tau + (1-q)w - \eta$, while her payoff from making an offer $\tau > k$ is w. A monitor with wage w and type η such that $\eta < k - qw$ finds it optimal to make an offer $\tau = k$, and a monitor with wage w and type η such that $\eta \geq k - qw$ finds it optimal to make offer $\tau > k$.

Finally, when the agent is non-corrupt, it is optimal for the monitor to send a truthful message $m = 0$ if there is no agreement at the collusion stage. Therefore, a non-corrupt agent is not willing to pay a bribe higher than 0 at the collusion stage. In this case, a monitor who acts as proposer demands a bribe $\tau = 0$ and sends a truthful message. ■

Proof of Proposition 2. The agent's payoff from taking action $c = 1$ is

$$\begin{aligned} U_A(\pi_A) &= \pi_A - k + \lambda \max_{\tau \in [0,\pi_A]} (k-\tau)\text{prob}(qw + \eta < \tau) \\ &= \pi_A - k + \lambda \max_{\tau \in [0,\pi_A]} (k-\tau)\mathbb{E}_{F_w}[F_\eta(\tau - qw)]. \end{aligned}$$

Consider first the case in which F_η is strictly concave $[0,k]$. Let τ_0 be the highest solution to the optimal bribe problem under a deterministic wage w_0 (i.e., $\max_\tau (k-\tau)F_\eta(\tau - qw_0)$) and note that $\tau_0 > qw_0$. Let F_w be a random wage distribution with $\mathbb{E}_{F_w}[w] = w_0$ and support $[w_0-\gamma, w_0+\gamma]$, with $\gamma > 0$ small enough such that $\tau_0 > q(w_0+\gamma)$. Let $F_w^\epsilon = (1-\epsilon)\mathbf{1}_{w=w_0}+\epsilon F_w$; i.e., policy F_w^ϵ is the mixture between a deterministic wage w_0 and policy F_w. Since F_η is strictly concave over $[0,k]$, $\mathbb{E}_{F_w^\epsilon}[F_\eta(\tau - qw)] < F_\eta(\tau - qw_0)$ for all τ close to τ_0. For each $\epsilon \in [0,1]$, let τ_ϵ be the highest solution to $\max_\tau (k-\tau)\mathbb{E}_{F_w^\epsilon}[F_\eta(\tau - qw)]$. Since τ_ϵ is close to τ_0 for ϵ small, it follows that for ϵ small the expected payoff that an agent gets from being corrupt under policy F_w^ϵ is strictly smaller than under the deterministic wage w_0.

Consider next the case in which F_η is strictly convex over $[0,k]$. Note that for any random wage distribution F_w with $\mathbb{E}_{F_w}[w] = w_0$, $F_\eta(\cdot)$ is convex over the support of $\tau - qw$ for all $\tau \in [0, \pi_A]$. Therefore, in this case the agent's payoff from being corrupt under any random wage distribution with mean w_0 is larger than under the deterministic policy w_0. ■

Proof of Proposition 3. For $\Delta > 0$, consider the random wage $\tilde{w}_\epsilon$ defined by

$$\tilde{w}_\epsilon = \begin{cases} w_0 - \epsilon & \text{with proba} \quad \frac{\Delta}{\Delta+\epsilon} \\ w_0 + \Delta & \text{with proba} \quad \frac{\epsilon}{\Delta+\epsilon}. \end{cases}$$

The expected payoff of a corrupt agent under random wage $\tilde{w}_\epsilon$ is

$$U_A(\pi_A|\tilde{w}_\epsilon) = \pi_A - k + \lambda \max_\tau (k-\tau)\text{prob}_{\tilde{w}_\epsilon}(qw + \eta < \tau).$$

By the Envelope Theorem,

$$\frac{\partial U_A(\pi_A|\tilde{w}_\epsilon)}{\partial \epsilon}_{|\epsilon=0} = (k-\tau_0)\left[-\frac{1}{\Delta}\text{prob}(qw_0 + \eta < \tau_0) + \frac{1}{\Delta}\text{prob}(q[w_0+\Delta] + \eta < \tau_0) + qf_\eta(\tau_0 - qw_0)\right].$$

Bribe τ_0, which solves $\max_\tau (k-\tau)\text{prob}(qw_0 + \eta < \tau)$, must be interior and therefore satisfies the first order condition

$$(k-\tau_0)f_\eta(\tau_0-qw_0)-\text{prob}(qw_0+\eta<\tau_0)=0\Rightarrow f_\eta(\tau_0-qw_0)=\frac{\text{prob}(qw_0+\eta<\tau_0)}{k-\tau_0}.$$

Setting $\Delta\equiv\tau_0/q-w_0$, we obtain that

$$\frac{\partial U_A(\pi_A|\tilde{w}_\epsilon)}{\partial\epsilon}_{\Big|\epsilon=0}=q(k-\tau_0)\text{prob}(qw_0+\eta<\tau_0)\left[-\frac{1}{\tau_0-qw_0}+\frac{1}{k-\tau_0}\right]<0$$

where we used the fact that $\tau_0\leq\frac{1}{2}k\Rightarrow k-\tau_0>\tau_0-qw_0$.

Hence for ϵ small enough, using random wage distribution $\tilde{w}_\epsilon$ reduces corruption compared to deterministic wage w_0. ■

Proof of Proposition 4. The payoff that an agent gets from being corrupt is $U_A(\pi_A)=\pi_A-k+\lambda\max_{\tau\in[0,\pi_A]}(k-\tau)\text{prob}(qw+\eta<\tau)$. For any wage schedule F_w, this payoff is maximized when $\lambda=1$ and when F_η is such that $F_\eta(0)=1$; that is, the worse case environment for the principal is that of Section 2.

By Proposition 1, in the worse case environment the cost minimizing distribution that induces an agent with private benefit π_A to take action $c=0$ is $F_w^{\mathsf{bmk}}=\frac{k-\pi_A}{k-qw}$. When the principal has a budget constraint w_0, the optimal wage schedule under the worse case environment is $F_w^{\mathsf{bmk}}=\frac{k-\overline{\pi}_A^0}{k-qw}$, where $\overline{\pi}_A^0$ is such that $W^{\mathsf{bmk}}(\overline{\pi}_A^0)=w_0$. ■

We now turn to the proof of Proposition 5, and begin with a few preliminary lemmas. It is useful to note that, for any wage schedule F_w, $\text{prob}(\eta+qw<\tau)=\int_0^{\frac{\tau}{q}}F_\eta(\tau-wq)dF_w(w)$. The incentive constraint (4) can then be written as: for all $\tau\in[0,\pi_A]$,

$$\int_0^{\frac{\tau}{q}}F_\eta(\tau-wq)dF_w(w)\leq\frac{k-\pi_A}{\lambda(k-\tau)}. \tag{14}$$

Note that, for an agent with type π_A and for any wage distribution F_w, (14) is satisfied for all $\tau \geq \overline{\tau} = \frac{\pi_A-(1-\lambda)k}{\lambda}$. Therefore, a principal who wants to incentivize agents with type $\pi'_A \leq \pi_A$ to be non-corrupt will never find it optimal to pay wages larger than $\frac{\overline{\tau}}{q}$. Based on this observation, when looking for the optimal distribution we can focus on c.d.f.s F_w such that $F_w(\overline{\tau}/q) = 1$.

Lemma B.1. *Suppose F_η is concave over $[0,k]$. If the distribution F_w satisfies (14) for all $\tau \in [0,\overline{\tau}]$ and $F_w(0) < m_0$, there exists a distribution $\tilde{F}_w$ which also satisfies (14) for all τ such that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$.*

Proof of Lemma B.1. Let F_w be a wage schedule that satisfies (14) for all $\tau \in [0,\overline{\tau}]$ with $F_w(0) < m_0$. Suppose first that F_w is such that (14) is satisfied with slack for all $\tau \in [0,\overline{\tau}]$. Let $\tilde{F}_w$ be a distribution such that $\tilde{F}_w(0) = F_w(0) + \gamma$ for some $\gamma > 0$, and such that $d\tilde{F}_w(w) = dF_w(w)$ for all $w \in (0,\tilde{w})$ (where $\tilde{w} = \inf\{w : \tilde{F}_w(w) = 1\}$). Clearly, $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. Moreover, since (14) is satisfied with slack for all τ under F_w, by choosing γ small we can guarantee that (14) is satisfied for all τ under $\tilde{F}_w$.

Suppose next that F_w is such that (14) binds for some offer τ. Let $\hat{\tau}$ be the lowest τ at which (14) binds, so that $\text{prob}(\eta + qw < \hat{\tau}) = \frac{k-\pi_A}{\lambda(k-\hat{\tau})}$. Since $F_w(0) < m_0$, it must be that $F_w(\frac{\hat{\tau}}{q}) > F_w(0)$: if $F_w(\frac{\hat{\tau}}{q}) = F_w(0)$, then $\text{prob}(\eta + qw < \hat{\tau}) = F_w(0)F_\eta(\hat{\tau}) = \frac{k-\pi_A}{\lambda(k-\hat{\tau})}$, which would imply that $F_w(0) = \frac{k-\pi_A}{F_\eta(\hat{\tau})\lambda(k-\hat{\tau})} \geq m_0$ (recall that m_0 is given by (5)).

We construct an alternative wage distribution $\hat{F}_w$ as follows. Fix $\gamma < F_w(\frac{\hat{\tau}}{q}) - F_w(0)$ and let $\hat{F}_w$ be such that: (i) $\hat{F}_w(0) = F_w(0) + \gamma$, (ii) $d\hat{F}_w(w) = 0$ for all $w \in (0, \frac{\hat{\tau}}{q})$, (iii) $d\hat{F}_w(w) = dF_w(w)$ for all $w \in (\frac{\hat{\tau}}{q}, \frac{\overline{\tau}}{q})$ and (iv) $d\hat{F}_w(\frac{\overline{\tau}}{q}) = 1 - \hat{F}_w(\frac{\overline{\tau}}{q}^-)$. Note that $\hat{F}_w$ is a transformation of F_w that shifts some of the mass that F_w has in $(0, \frac{\hat{\tau}}{q}]$ to 0 and the rest to $\frac{\overline{\tau}}{q}$. By choosing γ small we can guarantee that (14) is satisfied for all $\tau \in [0,\hat{\tau}]$ under $\hat{F}_w$.

We now show that (14) is also satisfied for all $\tau > \hat{\tau}$ under $\hat{F}_w$. Note first that for all

$\tau \geq \hat{\tau}$

$$\begin{aligned}\frac{\partial}{\partial \tau}\left(\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)dF_w(w)\right) &= \int_0^{\frac{\tau}{q}} f_\eta(\tau - wq)dF_w(w) > \\ \int_0^{\frac{\tau}{q}} f_\eta(\tau - wq)d\hat{F}_w(w) &= \frac{\partial}{\partial \tau}\left(\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w)\right),\end{aligned}$$

where the strict inequality follows since $\hat{F}_w$ puts more mass at 0 and less mass over $[0, \frac{\hat{\tau}}{q}]$ than F_w and since f_η is decreasing. Note that $\frac{k-\pi_A}{\lambda(k-\hat{\tau})} = \int_0^{\frac{\hat{\tau}}{q}} F_\eta(\hat{\tau} - wq)dF_w(w) \geq \int_0^{\frac{\hat{\tau}}{q}} F_\eta(\hat{\tau} - wq)d\hat{F}_w(w)$, where the equality follows since (14) binds at $\hat{\tau}$ under F_w and the inequality follows since (14) is satisfied at $\hat{\tau}$ under $\hat{F}_w$. Since (14) is satisfied for all τ under F_w, it follows that $\frac{k-\pi_A}{\lambda(k-\tau)} \geq \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)dF_w(w) > \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w)$ for all $\tau \in (\hat{\tau}, \overline{\tau})$; that is, (14) is satisfied with slack for all $\tau \in (\hat{\tau}, \overline{\tau})$ under $\hat{F}_w$.

For each $\varepsilon > 0$, let $\tilde{F}_\varepsilon$ be the wage schedule such that $d\tilde{F}_\varepsilon(w) = d\hat{F}_w(w)$ for all $w \notin \{\frac{\overline{\tau}-\varepsilon}{q}, \frac{\overline{\tau}}{q}\}$ and such that $d\tilde{F}_\varepsilon(\frac{\overline{\tau}-\varepsilon}{q}) = d\hat{F}_w(\frac{\overline{\tau}-\varepsilon}{q}) + d\hat{F}_w(\frac{\overline{\tau}}{q})$; i.e., $\tilde{F}_\varepsilon$ puts all the mass that $\hat{F}_w$ has on $\frac{\overline{\tau}}{q}$ at $\frac{\overline{\tau}-\varepsilon}{q}$. Note that, for all $\tau \leq \overline{\tau} - \varepsilon$, $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\tilde{F}_\varepsilon(w) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w) \leq \frac{k-\pi_A}{\lambda(k-\tau)}$; that is, for all $\varepsilon > 0$, (14) is satisfied for all $\tau \leq \overline{\tau} - \varepsilon$ under $\tilde{F}_\varepsilon$.

Note further that, for all $\tau \in (\overline{\tau} - \varepsilon, \overline{\tau})$, $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\tilde{F}_\varepsilon(w) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w) + F_\eta(\tau - (\overline{\tau} - \varepsilon))d\hat{F}_w(\frac{\overline{\tau}}{q})$ is continuous and increasing in ε. Since (14) holds with slack for all $\tau \in (\hat{\tau}, \overline{\tau})$ under $\hat{F}_w$, for ε small enough (14) also holds under wage schedule $\tilde{F}_\varepsilon$. Let $\overline{\varepsilon} \equiv \sup\{\varepsilon : \text{(14) holds for all } \tau \in [0, \overline{\tau}] \text{ under } \tilde{F}_\varepsilon\}$ and let $\tilde{F}_w = \tilde{F}_{\overline{\varepsilon}}$. Note that there must exist $\tau' > \overline{\tau} - \overline{\varepsilon}$ such that (14) holds with equality at τ' under $\tilde{F}_w$; i.e., such that

$$\frac{k - \pi_A}{\lambda(k - \tau')} = \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq)d\tilde{F}_w(w) \geq \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq)dF_w(w), \tag{15}$$

where the inequality follows since (14) holds for all τ under F_w.

We now use (15) to show that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. The distribution $\tilde{F}_w$ is a transformation of F_w that shifts some of the mass that F_w has on $[0, \frac{\hat{\tau}}{q}]$ to 0 and some mass up to $\frac{\overline{\tau}-\overline{\varepsilon}}{q}$. Since

F_η is strictly concave, (15) implies that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$; otherwise, if $\mathbb{E}_{\tilde{F}_w}[w] \geq \mathbb{E}_{F_w}[w]$ then F_w would second-order stochastically dominate $\tilde{F}_w$ and so (15) would not hold. ■

Lemma B.2. *Suppose F_w is such that $F_w(0) = m_0$. If* (14) *is satisfied for all τ under F_w, then it must be that $F_w(w) = m_0$ for all $w \in [0, \frac{\tau_0}{q}]$.*

Proof of Lemma B.2. Suppose by contradiction that $F_w(\frac{\tau_0}{q}) > F_w(0) = m_0$. Note then that $\int_0^{\frac{\tau_0}{q}} F_\eta(\tau_0 - wq) dF_w(w) > m_0 F_\eta(\tau_0) = \frac{k-\pi_A}{\lambda(k-\tau_0)}$, and so (14) does not hold at $\tau = \tau_0$. ■

Lemma B.3. *Suppose F_η is concave over $[0,k]$. Let F_w be a distribution with $F_w(w) = m_0$ for all $w \in [0, \frac{\tau_0}{q}]$ that satisfies* (14) *for all τ. If F_w is such that* (14) *doesn't hold with equality for all $\tau \in [\tau_0, \overline{\tau}]$ such that $F_w(\frac{\tau}{q}) < 1$, there exists a distribution $\tilde{F}_w$ which also satisfies* (14) *for all τ such that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$.*

Proof of Lemma B.3. Suppose that there is an interval (τ_1, τ_2) such that (14) is satisfied with slack for all $\tau \in (\tau_1, \tau_2)$ under F_w, with $\tau_1 \geq \tau_0$ and $F_w(\frac{\tau}{q}) < 1$ for all $\tau \in (\tau_1, \tau_2)$. There are two possibilities: (i) (14) does not bind for all $\tau > \tau_1$, or (ii) (14) binds at some $\hat{\tau} \geq \tau_2$. Consider first case (i) and let $\overline{w} = \inf\{w : F_w(w) = 1\}$ be the highest wage in the support of F_w. Fix $\gamma > 0$ and let $\tilde{F}_w$ be a wage distribution with $\tilde{F}(w) = F_w(w)$ for all $w < \overline{w} - \gamma$, and $\tilde{F}(\overline{w} - \gamma) = 1$. Clearly, $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. Since (14) is satisfied with slack for all $\tau > \tau_1$ under policy F_w, for γ small enough (14) is also satisfied for all τ under $\tilde{F}_w$.

Consider next case (ii). Without loss of generality, assume that (14) binds at τ_2. Fix $\gamma > 0$ and $\hat{\tau} \in (\tau_1, \tau_2)$ such that $\gamma(\hat{\tau} - \tau_1) < F_w(\frac{\tau_2}{q}) - F_w(\frac{\tau_1}{q})$. Let $\hat{F}_w$ be a wage distribution such that: (i) $\hat{F}_w(\frac{\tau}{q}) = F_w(\frac{\tau}{q})$ for all $\tau \leq \tau_1$, (ii) $\hat{F}_w(\frac{\tau}{q}) = F_w(\frac{\tau}{q}) + \gamma(\tau - \tau_1)$ for all $\tau \in (\tau_1, \hat{\tau})$, $d\hat{F}_w(\frac{\tau}{q}) = 0$ for all $\tau \in [\hat{\tau}, \tau_2)$, (iii) $d\hat{F}_w(\frac{\tau}{q}) = dF_w(\frac{\tau}{q})$ for all $\tau \in [\tau_2, \overline{\tau})$, and (iv) $d\hat{F}_w(\frac{\overline{\tau}}{q}) = 1 - \hat{F}_w(\frac{\overline{\tau}}{q}^-)$. Note that $\hat{F}_w$ is a transformation of F_w that shifts some of the mass that F_w has over $[\frac{\tau_1}{q}, \frac{\tau_2}{q}]$ to $[\frac{\tau_1}{q}, \frac{\hat{\tau}}{q}]$ and shifts the rest of this mass to $\frac{\overline{\tau}}{q}$. Since (14) is slack over (τ_1, τ_2)

under F_w, there exists γ and $\hat{\tau} \in (\tau_1, \tau_2)$ such that (14) is satisfied for all $(\tau_1, \tau_2]$ under $\hat{F}_w$. Moreover, since $\hat{F}_w(w) = F_w(w)$ for all $w \leq \frac{\tau_1}{q}$, (14) is satisfied for all $\tau \leq \tau_1$ under $\hat{F}_w$.

We now show that (14) also holds for all $\tau > \tau_2$ under $\hat{F}_w$. Note first that for all $\tau \geq \tau_2$

$$\begin{aligned}\frac{\partial}{\partial \tau}\left(\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)dF_w(w)\right) &= \int_0^{\frac{\tau}{q}} f_\eta(\tau - wq)dF_w(w) > \\ \int_0^{\frac{\tau}{q}} f_\eta(\tau - wq)d\hat{F}_w(w) &= \frac{\partial}{\partial \tau}\left(\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w)\right),\end{aligned}$$

where the strict inequality follows since $\hat{F}_w$ puts more mass on $[\frac{\tau_1}{q}, \frac{\hat{\tau}}{q}]$ but less mass over $[\frac{\tau_1}{q}, \frac{\tau_2}{q}]$ than F_w and since f_η is decreasing. Note that $\frac{k-\pi_A}{\lambda(k-\tau_2)} = \int_0^{\frac{\tau_2}{q}} F_\eta(\tau_2 - wq)dF_w(w) \geq \int_0^{\frac{\tau_2}{q}} F_\eta(\tau_2 - wq)d\hat{F}_w(w)$, where the equality follows since (14) binds at τ_2 under F_w and the inequality follows since (14) is satisfied at τ_2 under $\hat{F}_w$. Since (14) is satisfied for all τ under F_w, it follows that $\frac{k-\pi_A}{\lambda(k-\tau)} \geq \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)dF_w(w) > \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w)$ for all $\tau \in (\tau_2, \overline{\tau})$; that is, (14) is satisfied with slack for all $\tau \in (\tau_2, \overline{\tau})$ under $\hat{F}_w$.

The rest of the proof uses the same arguments as the proof of Lemma B.1. For each $\varepsilon > 0$, let $\tilde{F}_\varepsilon$ be such that $d\tilde{F}_\varepsilon(w) = d\hat{F}_w(w)$ for all $w \notin \{\frac{\overline{\tau}-\varepsilon}{q}, \frac{\overline{\tau}}{q}\}$ and such that $d\tilde{F}_\varepsilon(\frac{\overline{\tau}-\varepsilon}{q}) = d\hat{F}_w(\frac{\overline{\tau}-\varepsilon}{q}) + d\hat{F}_w(\frac{\overline{\tau}}{q})$. Note that $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\tilde{F}_\varepsilon(w) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w)$ for all $\tau \leq \overline{\tau} - \varepsilon$. Therefore, for all $\varepsilon > 0$, (14) holds for all $\tau \leq \overline{\tau} - \varepsilon$ under $\tilde{F}_\varepsilon$.

Let $\overline{\varepsilon} \equiv \sup\{\varepsilon : (14) \text{ holds for all } \tau \in [0, \overline{\tau}] \text{ under } \tilde{F}_\varepsilon\}$ and let $\tilde{F}_w = \tilde{F}_{\overline{\varepsilon}}$. Since $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\tilde{F}_\varepsilon(w)$ is continuous and increasing in ε for all $\tau > \overline{\tau} - \varepsilon$, there must exist $\tau' > \overline{\tau} - \overline{\varepsilon}$ such that (14) holds with equality at τ' under $\tilde{F}_w$; that is, such that

$$\frac{k - \pi_A}{\lambda(k - \tau')} = \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq)d\tilde{F}_w(w) \geq \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq)dF_w(w), \tag{16}$$

where the inequality follows since (14) holds for all τ under F_w. The distribution $\tilde{F}_w$ is a transformation of F_w that shifts some of the mass that F_w has on $[\frac{\tau_1}{q}, \frac{\tau_2}{q}]$ to $[\frac{\tau_1}{q}, \frac{\tau'}{q}]$ and the rest to $\frac{\overline{\tau}-\overline{\varepsilon}}{q}$. Since F_η is strictly concave, (16) implies that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$; otherwise, if

$\mathbb{E}_{\tilde{F}_w}[w] \geq \mathbb{E}_{F_w}[w]$ then F_w would second-order stochastically dominate $\tilde{F}_w$ and (16) would not hold. ■

Proof of Proposition 5. Let F_w^* be the optimal wage distribution. By Lemmas B.1 and B.2, $F_w^*(w) = m_0$ for all $w \in [0, \frac{\tau_0}{q}]$. By Lemma B.3, under F_w^* the constraint (4) holds with equality for all $\tau \in [\tau_0, \overline{\tau}]$ such that $F_w^*(\frac{\tau}{q}) < 1$; that is, for all τ in this range

$$H(\tau) \equiv \frac{k - \pi_A}{\lambda(k-\tau)} - \int_0^{\frac{\tau}{q}} F_\eta(\tau - \hat{w}q) dF_w^*(\hat{w}) = \frac{k - \pi_A}{\lambda(k-\tau)} - q\int_0^{\frac{\tau}{q}} f_\eta(\tau - \hat{w}q) F_w^*(\hat{w}) d\hat{w} = 0.$$

Therefore, for all $\tau \in [\tau_0, \overline{\tau}]$ such that $F_w^*(\frac{\tau}{q}) < 1$,

$$H'(\tau) = \frac{k - \pi_A}{\lambda(k-\tau)^2} - f_\eta(0) F_w^*\left(\frac{\tau}{q}\right) - q\int_0^{\frac{\tau}{q}} f'_\eta(\tau - q\hat{w}) F_w^*(\hat{w}) d\hat{w} = 0.$$

Using the change of variable $w = \frac{\tau}{q}$, for all $w \in [\frac{\tau_0}{q}, \frac{\overline{\tau}}{q}]$ such that $F_w^*(w) < 1$,

$$\begin{aligned} F_w^*(w) &= \frac{1}{f_\eta(0)}\left(\frac{k-\pi_A}{\lambda(k-qw)^2} - q\int_0^w f'_\eta(qw - q\hat{w}) F_w^*(\hat{w}) d\hat{w}\right) \\ &= \frac{1}{f_\eta(0)}\left(\frac{k-\pi_A}{\lambda(k-qw)^2} - \int_0^{qw} f'_\eta(\hat{\eta}) F_w^*\left(w - \frac{\hat{\eta}}{q}\right) d\eta\right). \end{aligned}$$

It follows that the optimal distribution F_w^* is the solution to $F_w^* = \Phi(F_w^*)$, where $\Phi(\cdot)$ is defined in (6).

Let F, G be two cdfs and let $\|\cdot\|$ denote the sup norm. Note that, for all $w \notin (\frac{\tau_0}{q}, \frac{\overline{\tau}}{q})$,

$|\Phi(F)(w) - \Phi(G)(w)| = 0$. On the other hand, for all $w \in (\frac{\tau_0}{q}, \frac{\overline{\tau}}{q})$,

$$\begin{aligned}
|\Phi(F)(w) - \Phi(G)(w)| &\leq \left| \frac{-1}{f_\eta(0)} \int_0^{qw} f_\eta'(\eta) \left(F\left(w - \frac{\eta}{q}\right) - G\left(w - \frac{\eta}{q}\right)\right) d\eta \right| \\
&\leq \|F - G\| \left| \frac{-1}{f_\eta(0)} \int_0^{qw} f_\eta'(\eta) d\eta \right| \\
&= \|F - G\| \frac{f_\eta(0) - f_\eta(qw)}{f_\eta(0)} \\
&\leq \|F - G\| \frac{f_\eta(0) - f_\eta(\overline{\tau})}{f_\eta(0)},
\end{aligned}$$

where the last inequality follows since f_η is decreasing. Note that $\overline{\tau} = \frac{\pi_A - (1-\lambda)k}{\lambda} < k$. Since $f_\eta(\cdot)$ is strictly positive for all $w \in [0, k]$, $d \equiv \frac{f_\eta(0) - f_\eta(\overline{\tau})}{f_\eta(0)} < 1$. It follows that $\|\Phi(F) - \Phi(G)\| \leq d\|F - G\|$, so Φ is a contraction mapping of modulus $d < 1$. ■

B.3 Proofs for Section 4

Proof of Fact 4. The proof is by example. We proceed case by case and assume throughout that $\lambda = 1$. Denote by $\overline{w}$ and $\underline{w}$ the maximum and minimum values in the support of F_w^1. Note that $w_0 \in (\underline{w}, \overline{w})$.

We first show that $\overline{R}_0 < \overline{R}_1$ can be consistent with $\overline{C}_0 < \overline{C}_1$. Consider the case where $k = qw_0$, F_{π_A} is a mass point at $k - \epsilon$ with $\epsilon > 0$, and F_η a mass point at 0. For any $\epsilon > 0$, $\overline{R}_0 = \overline{C}_0 = 0$. For $\epsilon > 0$ small enough $F_w^1(w_0 - \epsilon) > 0$, which implies that for ϵ small enough,

$$\max_\tau (k - \tau) \text{prob}_{F_w^1}(qw < \tau) > \epsilon.$$

Hence for $\epsilon > 0$ small enough, $\overline{C}_1 = 1$. Furthermore, for $\epsilon > 0$ small enough, $F_w^1(w_0 + \epsilon) < 1$, which implies that $\overline{R}_1 > 0$ since the agent never offers a bribe $\tau \geq k = qw_0$.

Let us show that $\overline{R}_0 < \overline{R}_1$ can be consistent with $\overline{C}_0 > \overline{C}_1$. Set F_{π_A} with full support over $[0,k]$, and

$$\eta = \begin{cases} \overline{\eta} & \text{with proba} \quad p \\ 0 & \text{with proba} \quad 1-p \end{cases}$$

with both $\overline{\eta} \leq \epsilon$ and $p \leq \epsilon$. For k large enough and $\epsilon > 0$ small enough, it is immediate that

$$\max_{\tau}(k-\tau)\text{prob}_{F_w^1}(qw+\eta<\tau) < \max_{\tau}(k-\tau)\text{prob}(qw_0+\eta<\tau)$$

since as k grows large, it is optimal for the agent to offer bribes respectively converging to $\overline{w}$ and w_0, and $\overline{w} > w_0$. This implies that $\overline{C}_0 > \overline{C}_1$. Let us now show that we can set $\overline{\eta}$ and p so that $\overline{R}_0 < \overline{R}_1$. A necessary and sufficient condition to obtain $\overline{R}_0 = 0$

$$k - qw_0 - \overline{\eta} > (k-qw_0)(1-p) \iff k - qw_0 > \frac{\overline{\eta}}{p}. \tag{17}$$

This condition expresses that it is optimal for the agent to offer a bribe $\tau = qw_0 + \overline{\eta}$ rather than $\tau = qw_0$. Similarly, under F_w^1. a sufficient condition to ensure that $\overline{R}_1 > 0$ is that the agent prefer offering a bribe $\tau = q\overline{w}$ over bribe $\tau = q\overline{w} + \overline{\eta}$. A sufficient condition for this is that

$$k - q\overline{w} - \overline{\eta} < (k-q\overline{w})(1-p) \iff k - q\overline{w} < \frac{\overline{\eta}}{p}. \tag{18}$$

Since $\overline{w} > w_0$, it is immediate that for any ϵ, one can find values $p, \overline{\eta} < \epsilon$, such that conditions (17) and (18) hold simultaneously. For such values, $\overline{R}_1 > \overline{R}_0 = 0$, which yields the desired result.

We now show that $\overline{R}_0 > \overline{R}_1$ can be consistent with $\overline{C}_0 > \overline{C}_1$. Set

$$\eta = \begin{cases} \overline{\eta} & \text{with proba} \quad p \\ 0 & \text{with proba} \quad 1-p \end{cases}$$

with both $\overline{\eta} \leq \epsilon$ and $p \leq \epsilon$. For k large enough and $\epsilon > 0$ small enough, we have that

$$\max_{\tau}(k-\tau)\text{prob}_{F_w^1}(qw+\eta<\tau) < \max_{\tau}(k-\tau)\text{prob}(qw_0+\eta<\tau).$$

Set F_{π_A} as a point mass at a value π_A such that

$$\pi_A - k + \max_{\tau}(k-\tau)\text{prob}_{F_w^1}(qw+\eta<\tau) < 0 < \pi_A - k + \max_{\tau}(k-\tau)\text{prob}(qw_0+\eta<\tau)$$

for all ϵ small enough. This implies that $\overline{C}_0 = 1 > \overline{C}_1 = 0$. In turn we obtain that $\overline{R}_1 = 0$. Finally, by choosing p and $\overline{\eta}$ such that (17) does not hold, one can ensure that $\overline{R}_0 > 0$.

Finally, we show that $\overline{R}_0 > \overline{R}_1$ can be consistent with $\overline{C}_0 < \overline{C}_1$. Set $\eta = 0$, $k = qw_0 - \frac{1}{2}\epsilon$ and

$$\pi_A = \begin{cases} k+\epsilon & \text{with proba} \quad p \\ k & \text{with proba} \quad 1-p. \end{cases}$$

It is immediate that $\overline{C}_0 = p$ and $\overline{R}_0 = p$. Furthermore, since $\max_\tau(k-\tau)\text{prob}_{F_w^1}(qw+\eta<\tau)$ is strictly positive and bounded away from 0 for ϵ small enough, it follows that for ϵ small enough $\overline{C}_1 = 1$ and $\overline{R}_1 < 1$. For p large enough, $\overline{R}_0 > \overline{R}_1$. This concludes the proof. ■

Proof of Proposition 6. Under wage schedule F_w^ϵ, the agent's payoff from action $c = 1$ is

$$U_A^\epsilon(\pi_A) = \pi_A - k + \lambda \max_{\tau}(k-\tau)\left[(1-\epsilon)\text{prob}_{F_w^0}(qw+\eta<\tau) + \epsilon\text{prob}_{F_w^1}(qw+\eta<\tau)\right].$$

Let τ_ϵ be the offer that maximizes this expression. By the Envelope Theorem,

$$\frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon} \quad = \quad (k-\tau_\epsilon)\left[\text{prob}_{F_w^1}(qw+\eta<\tau) - \text{prob}_{F_w^0}(qw+\eta<\tau)\right].$$

Proposition 6 follows since $\mathsf{sgn}\left[\frac{\partial \overline{C}_\epsilon}{\partial \epsilon}\right] = \mathsf{sgn}\left[\frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}\right]$ and since $\mathsf{sgn}\left[\text{prob}_{F_w^1}(qw+\eta<\tau) - \text{prob}_{F_w^0}(qw+\eta<\tau)\right] = \mathsf{sgn}\left[R^\epsilon(X=0) - R^\epsilon(X=1)\right]$. ■

References

Ashraf, N., J. Berry, and J. M. Shapiro (2010): "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *The American economic review*, 100, 2383–2413.

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

Banerjee, A., S. Mullainathan, and R. Hanna (2013): *Corruption*, Princeton University Press.

Becker, G. S. and G. J. Stigler (1974): "Law enforcement, malfeasance, and compensation of enforces," *J. Legal Stud.*, 3, 1.

Berry, J., G. Fischer, and R. Guiteras (2012): "Eliciting and utilizing willingness to pay: evidence from field trials in Northern Ghana," *Unpublished manuscript*.

Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan (2007): "Obtaining a driver's license in India: an experimental approach to studying corruption," *The Quarterly Journal of Economics*, 122, 1639–1676.

Brooks, B. (2014): "Surveying and selling: Belief and surplus extraction in auctions," .

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. (2013): "Calibrated incentive contracts," *Econometrica*, 81, 1935–1971.

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

Chassang, S., G. Padró i Miquel, and E. Snowberg (2012): "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 102, 1279–1309.

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

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2013): "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India*," *The Quarterly Journal of Economics*, 128, 1499–1545.

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.

Fisman, R. and S.-J. Wei (2004): "Tax Rates and Tax Evasion: Evidence from Missing Imports in China," *Journal of Political Economy*, 112.

Frankel, A. (2014): "Aligned delegation," *The American Economic Review*, 104, 66–83.

Hartline, J. D. and T. Roughgarden (2008): "Optimal Mechanism Design and Money Burning," in *Symposium on Theory Of Computing (STOC)*, 75–84.

Hurwicz, L. and L. Shapiro (1978): "Incentive structures maximizing residual gain under incomplete information," *The Bell Journal of Economics*, 9, 180–191.

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

Karlan, D. and J. Zinman (2009): "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment," *Econometrica*, 77, 1993–2008.

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.

Madarász, K. and A. Prat (2014): "Screening with an Approximate Type Space," *Working Paper, London School of Economics.*

Maskin, E. (1999): "Nash equilibrium and welfare optimality*," *The Review of Economic Studies*, 66, 23–38.

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

Myerson, R. B. (1986): "Multistage games with communication," *Econometrica: Journal of the Econometric Society*, 323–358.

Myerson, R. B. and M. A. Satterthwaite (1983): "Efficient mechanisms for bilateral trading," *Journal of economic theory*, 29, 265–281.

Olken, B. A. (2007): "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 115.

Olken, B. A. and R. Pande (2012): "Corruption in Developing Countries," *Annual Review of Economics*, 4, 479–509.

Prat, A. (2014): "Media Power," *Columbia University Working Paper.*

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.

Zitzewitz, E. (2012): "Forensic economics," *Journal of Economic Literature*, 50, 731–769.