Making Corruption Harder:

Asymmetric Information, Collusion, and Crime*

Juan Ortner
Boston University

Sylvain Chassang†
Princeton University

June 3, 2016

Abstract

We model the investigation of criminal activity as a principal-agent-monitor problem in which the agent can corrupt the monitor and side-contract to destroy evidence. Building on insights from Laffont and Martimort (1997) we study whether the principal can benefit from endogenously creating asymmetric information between the agent and the monitor. We show that the principal can potentially obtain large benefits from randomizing the incentives given to the monitor (and letting those serve as the monitor's private information), but that the optimality of random incentives depends on pre-existing patterns of private information. This motivates us to develop a data-driven framework for policy evaluation using only unverified report data.

KEYWORDS: monitoring, collusion, corruption, asymmetric information, random incentives, prior-free policy evaluation.

*An early version of this paper was circulated under the title "Making Collusion Hard: Asymmetric Information as a Counter-Corruption Measure." We are especially thankful to Gerard Padró i Miquel for many helpful and inspiring conversations. We are indebted to Yeon-Koo Che, Hugo Hopenhayn, Bart Lipman, Dilip Mookherjee, Stephen Morris, Andy Newman, Debraj Ray as well as seminar participants at Arizona State University, Caltech, the Canadian Economic Theory Conference 2015, Collegio Carlo Alberto, Columbia, the 13th Columbia/Duke/MIT/Northwestern IO Theory Conference, the Einaudi Institute, Iowa State, McGill, Minnesota, Rice, SITE, UT Austin, Warwick, and Wash. U. 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

Agents potentially engaging in criminal behavior can limit the effectiveness of the judiciary by corrupting monitors in charge of investigating them. This paper explores the idea that corruption can be reduced by introducing endogenous asymmetric information frictions between colluding parties. Building on seminal work by Laffont and Martimort (1997), we show that the cost of deterring crime can be reduced 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 efficiency gains are large in plausible settings, but in general, the optimality of random incentives depends on pre-existing patterns of asymmetric information, making policy design difficult. We address this problem by providing a data-driven framework for prior-free policy evaluation: although aggregate reports by monitors cannot be naïvely used to measure actual criminal activity, we show how to evaluate policy changes using only unverified report data.

We study a game between three players — a principal, an agent, and a monitor — in which the agent chooses whether or not to engage in criminal behavior $c \in \{0, 1\}$. The behavior of the agent is not observed by the principal, but is observed by a monitor who submits report $m \in \{0, 1\}$. We think of this report as evidence leading to prosecution: report $m = 1$ triggers an exogenous judiciary process which imposes a cost k on criminal agents; report $m = 0$ (which involves suppression of evidence whenever $c = 1$) triggers no such process. Although the principal cannot observe the agent's behavior, she can detect misreporting $m \neq c$ with probability q. The principal's only policy control is the efficiency wage w provided to the monitor.

We allow for collusion between the agent and the monitor at the reporting stage (i.e. corruption). In particular, the monitor can destroy evidence (report message $m = 0$) incriminating a criminal 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. Exploiting the classic

insight that asymmetric information may prevent efficient trade and limits collusion (Myerson and Satterthwaite, 1983, Laffont and Martimort, 1997), 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 investigate. 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), organized crime and police officers (Punch, 2009), and so on. In these settings the principal cannot efficiently monitor agents directly, but may realistically be able to detect tempered evidence by scrutinizing accounts, performing random rechecks in person, or obtaining tips from informed parties. Alternatively, the principal may be able to detect misreporting if crime 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 cheap 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 criminal agents, 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, in excess of 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 random incentives depends on the convexity or concavity of the c.d.f. F_η of idiosyncratic costs η.

Finally, motivated by the fact that optimal policy depends crucially on fine details of the environment, we study the possibility of performing prior-free policy evaluations using reporting data from a population of agent-monitor pairs. We first show that aggregate reports of crime across different incentive schemes do not allow for reliable policy evaluation. Indeed, reports of crime depend on both underlying crime rates, and the monitors' decision to report crime or not. As a result, it is possible that a new incentive scheme decreases aggregate reports of crime, while in fact increasing underlying crime rates. Surprisingly, we are able to show that it is possible to perform prior-free local policy evaluations using conditional report data from a single policy (i.e. average reports of crime conditional on realized incentives). Somewhat counter-intuitively, a local policy change improves on a reference incentive scheme if it is associated with higher rates of reported crime.

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 between the monitor and the agent, 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 and the agent must incentivize her preferred report by committing to a retaliation strategy. Chassang and Padró i Miquel (2013) show that it is important for the principal to limit the information content of her own response to the monitor's reports. 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 incentive schemes may be very different over the short-run and the long-run: over time, agents will find ways to corrupt the investigators in charge of monitoring them. We explicitly take into account the possibility of collusion between agents and monitors and propose ways to reduce the costs it imposes on organizations. A second difficulty brought up by Olken and Pande (2012) is that reports of criminal behavior do not provide a reliable measure of underlying crime. Our structural model allows us to back-out measures of underlying crime using observed reports. 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 initiated by Tirole (1986). It is especially related to Laffont and Martimort (1997, 2000) and Che and Kim (2006, 2009), who emphasize the role of asymmetric information in limiting the extent of collusion.[2] Our contribution is two-fold. First, we show that the principal can potentially benefit from introducing endogenous asymmetric information through random incentives.[3] Second, as a step towards implementation, we show how to evaluate potential policy changes using only unverified reports. Also related is Baliga and Sjöström (1998), who suggest a distinct mechanism through which random wages (to the agent) may help reduce collusion. They 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. When that is the case, randomizing the agent's wages undermines her ability to

[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]For more on the large literature on collusion in mechanism design, see Felli and Villa-Boas (2000), Faure-Grimaud et al. (2003), Mookherjee and Tsumagari (2004), Burguet and Che (2004), Pavlov (2008), Celik (2009) or Che et al. (2013).

[3]This relates our paper to a recent literature that studies optimal design of information structures; see, for instance, Bergemann and Pesendorfer (2007), Kamenica and Gentzkow (2011), Bergemann et al. (2015), Condorelli and Szentes (2016).

commit to transfers.[4]

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.[5] 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.[6]

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 that make random incentives useful. Section 3 extends the analysis to environments with pre-existing asymmetric information, and shows that additional asymmetric information need not always be optimal. Section 4 proposes an approach to policy-evaluation relying on naturally occurring report data. Appendix A presents several extensions describing how our results extend in settings allowing for: more sophisticated contracting between the principal and monitor, multiple monitors, efficient incomplete-information bargaining between the monitor and agent, extortion from non-criminal agents, and repeated interaction. Appendix B provides an explicit characterization of maxmin and Bayesian-optimal incentive profiles. Proofs are collected in Appendix C unless mentioned otherwise.

[4]Basu (2011) and Basu et al. (2014) highlight the role of asymmetric punishments in reducing collusion and bribery.

[5]See Lazear (2006) for related results.

[6]Other papers have emphasized the role of random incentives. Rahman and Obara (2010) demonstrate that random messages can improve incentive provision in partnerships by allowing to identify innocent individuals. Jehiel (2012) shows that a principal may benefit from maintaining her agent uninformed about payoff relevant features of the environment, as this may induce higher effort at states at which she values effort most. In a multi-tasking setting, Ederer et al. (2013) show that random contracts may be effective in incentivizing the agent to take a balanced effort profile. In a monopoly pricing context, Calzolari and Pavan (2006a,b) 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.

2 A Simple Example

2.1 Framework

Players, actions, and payoffs. We consider a game with three players: a principal, e.g. an environmental protection agency (EPA); an agent, e.g. an industrial plant; and a monitor, e.g. an investigator employed by the EPA. The agent decides whether to engage in criminal behavior $c \in \{0, 1\}$, where crime $c = 1$ gives the agent a benefit $\pi_A > 0$, and comes at a cost $\pi_P < 0$ to the principal. For instance, the industrial plant may choose to dump hazardous materials rather than incur the cost of processing them. 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. We think of this report as evidence leading to prosecution: report $m = 1$ triggers a judiciary process that imposes an expected cost $k > \pi_A$ on criminal agents and an expected cost $k_0 \in [0, k]$ on non-criminal agents.[7] This judiciary process is exogenous and outside the control of the principal.[8]

While reports can be falsified (the monitor can always send either reports), we assume that the principal detects false reports $m \neq c$ with probability $q \in (0, 1)$, which makes reports partially verifiable. Detection may occur through several channels: accounting discrepancies, random rechecks, tips from informed parties. Criminal behavior may also have delayed but observable consequences, such as environmental pollution.

For most of the paper, we assume that the monitor is paid according to a fixed wage contract with wage w, and 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. We show in Appendix A how our results extend when the principal uses arbitrary

[7] Note that in the US, environmental pollution is indeed subject to criminal prosecution. The EPA maintains a database of criminal cases resulting from its investigations at http://www2.epa.gov/enforcement/summary-criminal-prosecutions.

The cost k_0 that the judiciary imposes on non-criminal agents does not play a role in our analysis, except when we consider arbitrary contracts and the possibility of extortion from non-criminal agents in Appendix A.

[8] Our results generalize to settings in which the principal can also contract with the agent.

contracts to compensate the monitor. The principal benefits from using more sophisticated contracts, since she can offer the monitor a large payment following report $m = 1$. However, our assumption that the monitor's messages are only partially verifiable limits the gains from using such compensation schemes: as the payment following report $m = 1$ grows large, the monitor has an incentive to report crime regardless of the agent's action.

As part of a possible side-contract the agent can make transfers $\tau \geq 0$ to the monitor, i.e. pay her a bribe. Corruption occurs when the monitor accepts to destroy evidence for a criminal agent (i.e. sends message $m = 0$ although $c = 1$). Note that crime rather than corruption is the behavior that the principal really cares about. Corruption undermines the effectiveness of institutions in charge of punishing crime.

Altogether, expected payoffs u_P, u_A, and u_M respectively accruing to the principal, the agent, and the monitor take the form:

$$\begin{aligned}
u_P &= \pi_P \times c \quad -\gamma_w \times w - \gamma_q \times q \\
u_A &= \pi_A \times c \qquad\qquad\qquad -[k \times c + k_0 \times (1-c)] \times m \quad -\tau \\
u_M &= w \qquad\qquad\qquad\qquad\quad -q \times w \times \mathbf{1}_{m \neq c} \qquad\qquad +\tau,
\end{aligned}$$

where γ_w denotes the efficiency cost of raising promised wages and γ_q captures the principal's cost of attention. When the principal is operating under budget or attention constraints, these costs can be interpreted as shadow prices. We assume for now that parameters π_A, k, k_0 and q are common knowledge. We relax this assumption in Section 4.

Note that the monitor's incentives for truthful reporting are captured by the expected loss from misreporting qw. For ease of exposition we treat the distribution of wages w as the principal's policy variable. However, note 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, if 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 chooses whether or not to engage in crime $c \in \{0, 1\}$;[9]
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.

We assume for now that the agent has all the bargaining power at the collusion stage, if it occurs. We show that our results extend under more general bargaining (Section 3, Appendix A), including in environments where the monitor may try to reveal her incentives to the agent. While the monitor would indeed benefit from revealing her incentives, Appendix A shows that it is difficult to do so in a credible way. Regardless of her type, the monitor would like to convince the agent that she has the incentives for truth-telling which lead to the highest bribe.

Our model admits a natural population interpretation in which distribution F_w captures wage heterogeneity in the population of monitors. We assume that the principal can commit to such a distribution, which is plausible when the principal operates under a fixed budget. Section 5 suggests credible ways to create asymmetric information without randomization, for instance by using time-varying or non-linear incentives.[10]

[9] We assume that the agent and monitor bargain after the agent decides to engage in criminal behavior. This is consistent with the idea that the monitor's ability to commit to reports is achieved by destroying incriminating evidence, or filling her report in front of the agent. In any case, we show in Appendix A that our main qualitative results do not change if bargaining occurs before the agent's decision: it is still optimal for the principal to use random incentives.

[10] We also emphasize that heterogeneity in incentives, rather than heterogeneity in wages is at the core of our argument. This could be achieved through heterogeneity in q, i.e. in the scrutiny monitors are under.

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

Motivation. Our framework is intended to capture the challenges facing public agencies that rely on monitors to assess the behavior of regulated agents. Besides environmental protection, prominent examples include labor safety regulation, tax collection, and tackling organized crime. In these cases, crime may respectively correspond to maintaining poor safety standards, fraudulent accounting, or extortion and smuggling. The monitor may commit not to report the agent by destroying, or simply by not collecting the evidence needed to initiate a judiciary process. Even if the monitor makes no report of crime, signals of misbehavior may be obtained by the principal after some delay: pollution or poor safety standards may lead to visible consequences (e.g. accidents, local contamination); civil society stakeholders may produce evidence of their own; aggrieved associates of the agent may volunteer incriminating information; and so on ...

2.2 The value of endogenous asymmetric information

We now characterize optimal policy in this simple setting. The following observation is useful.

Lemma 1. *Under* collusion*, the monitor will accept a bribe τ from a criminal agent if and only if $\tau > qw$.*[11] *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 $m = c$.*

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

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

Deterministic wages. We begin by computing the cost of keeping the agent non-criminal when the principal can use only deterministic wages.

Lemma 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-criminal at 0 cost.*

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

Proof. By Lemma 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-criminal at zero cost.

Consider next a setting with *collusion*. By Lemma 1, the monitor accepts a bribe τ from a criminal 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 $w = \frac{\pi_A}{q}$. ■

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). We show in Appendix A that 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.

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^* that induces the agent to be non-criminal is described by*

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

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

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

The proof of Proposition 1 is instructive.

Proof. A wage distribution F_w induces the agent to be non-criminal 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 $\tau \in [0, \pi_A]$, $F_w\left(\frac{\tau}{q}\right) \leq \frac{k - \pi_A}{k - \tau}$. Using the change in variable $w = \frac{\tau}{q}$, we obtain that wage distribution F_w induces the agent to be non-criminal if and only if,

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

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. ■

Further intuition for why random wages can improve on deterministic wages can be obtained by considering small perturbations around determinisic wage $\frac{\pi_A}{q}$. Wage $\frac{\pi_A}{q}$ deters crime since a criminal agent finds it optimal to offer bribe $\tau = \pi_A$, which absorbs all the potential profits from crime. Consider now setting a wage equal to $\frac{\pi_A}{q}$ with probability $1 - \epsilon$ and equal to zero otherwise. Since the cost k of prosecution is strictly higher than π_A, for $\epsilon > 0$ small enough, a criminal agent will still offer a bribe $\tau = \pi_A$. This lets the principal

deter crime at a lower expected cost of incentives.

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 than $\frac{\pi_A}{q}\frac{\pi_A}{k}$ for the optimal random incentive scheme. For instance, if the penalty for crime is greater than twice its benefits, 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.[12]

Are random incentives robustly optimal? Because the efficiency gains from using random incentives appear large in this simple example, we want to take seriously the possibility of field implementation. For this, we need to better assess the robustness of our results. We are able to show in Appendix A that relaxing the assumptions of efficiency wages and take-it-or-leave-it-bargaining does not overturn the optimality of random incentives (see also the discussion provided in Section 5). Pre-existing asymmetric information poses a more fundamental challenge to our findings.

The fact that complete information should overstate the value of random incentives is intuitive: under complete information, random incentives are the only private information allowing the monitor to extract rents. Section 3 shows that in fact, random incentives need not be helpful when the monitor incurs a privately observed cost for accepting bribes. The optimality of random incentives depends on the distribution of such costs, which makes policy recommendation difficult. Section 4 addresses this issue by showing it is possible to perform local policy evaluations using naturally occuring report data.

[12]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 $w = \pi_A/q$.

3 Pre-existing Asymmetric Information

3.1 Framework

We extend the model of Section 2 in three ways:

- 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;
- the agent's benefit π_A from crime is now private information to the agent, distributed according to c.d.f. F_{π_A} with density f_{π_A};
- 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 $1 - \lambda$.

Distributions F_η and F_{π_A} are naturally interpreted as sample distributions of types in a large population of monitors and agents. We maintain this population interpretation of the model throughout the rest of the paper.

Altogether, payoffs now take the form

$$\begin{array}{lllll} u_P & = \pi_P \times c & -\gamma_w \times w - \gamma_q \times q & & \\ u_A & = \pi_A \times c & & -[k \times c + k_0 \times (1-c)] \times m & -\tau \\ u_M & = w & & -[q \times w + \eta] \times \mathbf{1}_{m \neq c} & +\tau. \end{array}$$

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 and sending a false report, 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, k_0, λ and q are common-knowledge between the agent and the monitor.

The following extension of Lemma 1 holds.

Lemma 3. *If no agreement is reached at the collusion stage, a monitor's optimal continuation strategy is to send truthful reports $m = c$.*[13]

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

The agent accepts any offer $\tau \leq k$ when she is criminal.

An immediate implication is that non-criminal agents get a payoff equal to 0.[14]

Policy design under budget constraints. Given a distribution of wages F_w, a criminal agent of type π_A gets an expected payoff

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

An agent will engage in crime 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 of wages F_w, let us denote by $\overline{\pi}_A(F_w)$ the value of π_A for which an agent is indifferent between actions $c = 0$ and $c = 1$.

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 *crime-minimizing* wage schedule, given budget w_0. The overall optimum can then be obtained by optimizing over budget w_0. We are principally interested in this fixed-budget version of the principal's problem, which reflects the budget constraints that real-life institutions frequently operate under.

Our population interpretation of the model means that the principal can satisfy budget constraint $\mathbb{E}_{F_w}[w] = w_0$ exactly while using a non-degenerate distribution of wages. In

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

[14] To see why the monitor demands bribe $\tau \geq k$ when the agent is criminal, note that a monitor with type η and wage w obtains a payoff of $\tau + (1 - q)w - \eta$ by making an offer $\tau \leq k$ that the agent accepts, and obtains a payoff of w by making an offer $\tau > k$ that the agent rejects. Therefore, it is optimal for the monitor to demand $\tau = k$ if $k > qw + \eta$, and to demand $\tau > k$ if $k \leq qw + \eta$.

addition, fixed budgets support the principal's ability to commit to mixed strategies. Indeed, taking agent behavior as given, the principal is indifferent over distributions $\tilde{F}_w$ satisfying $\mathbb{E}_{\tilde{F}_w}[w] = w_0$.

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 crime-minimizing wage profile under any budget $w_0 > 0$ is random.*

(ii) Whenever F_η is strictly convex over the range $[0, k]$, the crime-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 $\tau - qw$, a criminal agent is effectively risk-loving and she obtains a higher payoff from a random wage schedule than from a deterministic one with the same expectation. Inversely, if F_η is strictly concave over the support of $\tau - qw$, a criminal agent is effectively risk-averse and her payoff from a random wage schedule is smaller than her payoff from a deterministic one with the same expectation.

If F_η 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 criminal 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 crime-minimizing policy given budget w_0 is random.*

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. The proof exploits the fact that c.d.f. F_η cannot be convex over arbitrarily large ranges of values.[15] The assumption that $\tau_0 \leq \frac{k}{2}$ lets us exploit non-convexities of F_η around w_0 to construct random wage schedules that improve on fixed wages.

Because adding further asymmetric information does not necessarily improve incentive provision, correct policy design must depend on the restrictions, subjective or objective, that the principal can impose on the environment. We refer the reader to Appendix B for a characterization of Bayesian-optimal policies in well-behaved cases. However, specifying beliefs is often difficult for principals, which makes actual implementation difficult. To address the issue, we show in the next section that it is possible to perform prior-free policy evaluations using naturally occurring report data. Our result on policy evaluation allows a principal to search for optimal policies even if she has little knowledge about the environment.

4 Prior-free Policy Evaluation

We now show that it is possible to evaluate potential local policy changes using only the reporting data occurring under existing policies. Our 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 criminal agents, the likelihood q of detection, and bargaining power λ need not

[15]Note that $\tau_0 \leq \frac{k}{2}$ implies that F_η is not convex over $[0,k]$. Indeed, optimal bribe τ_0 must satisfy the first-order condition $f_\eta(\tau_0-qw_0)(k-\tau_0)=F_\eta(\tau_0-qw_0)$. If F_η was convex over $[0,k]$, then $F_\eta(\tau_0-qw_0)\leq f_\eta(\tau_0-qw_0)(\tau_0-qw_0)<f_\eta(\tau_0-qw_0)(k-\tau_0)$, where the last inequality follows since $\tau_0\leq\frac{k}{2}$ and $w_0>0$.

be known to the observer). We also emphasize that no experimental variation is needed for local policy evaluation: it is sufficient to obtain reporting data at a single full-support wage policy. In particular, equilibrium reporting data at alternative policies is not required. This implies that evaluating policy changes need not require long costly experiments.

Note that implementing random incentive schemes is not implausible (see Section 5 for further discussion). For instance, Khan et al. (2014) vary the piece-rates given to different tax-collectors, resulting in a distribution of incentives. Our results show that reporting data from a single such implementation lets us evaluate whether random incentives would reduce the cost of containing crime.

Naïve inference fails. We begin by showing that a naïve use of reporting data from natural policy experiments fails to identify the true effect of policy changes. Given budget w_0, consider two policies F_w^0, F_w^1 such that $\mathsf{supp}\, F_w^0 = \{w_0\}$ (i.e. F_w^0 is a reference fixed-wage policy), and $\mathbb{E}_{F_w^1}[w] = w_0$. Imagine that policies F_w^ϵ for $\epsilon \in \{0,1\}$ are 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 crime 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 crime, and by $\overline{C}_\epsilon = \widehat{\mathbb{E}}[c^\epsilon]$ the proportion of criminal agents.

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

Regardless of the ranking of reports, i.e. 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 effect of random incentives on crime. Indeed, reports of crime depend on both underlying rates of crime, and the monitors' decisions to report it. A change in incentive patterns from F_w^0 to F_w^1 changes

both the agents' decision to engage in crime and their bribing behavior. As a result, changes in aggregate reports from $\overline{R}_0$ to $\overline{R}_1$ do not always match changes in underlying crime.

Local policy evaluation. We now show that an appropriate use of report data from policies with non-trivial support can be used to evaluate *local* policy changes.

Take as given a distribution of wages with cdf F_w^0 and density f_w^0. Denote by $\mathcal{P}_0$ the set of alternative policies f_w^1 satisfying

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

When f_w^0 has full support, the set of policies $f_w^1 \in \mathcal{P}_0$ is simply the set of policies with the same wage bill as f_w^0.

For any alternative policy f_w^1, construct the mixture $f_w^\epsilon = (1-\epsilon)f_w^0 + \epsilon f_w^1$ and let $\widehat{\mathbb{E}}[c^\epsilon | f_w^\epsilon]$ be the proportion of criminal agents under policy f_w^ϵ. Denote by $\nabla_{f_w^1}\overline{C}$ the gradient of equilibrium crime in policy direction f_w^1:

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

We are interested in evaluating the gradient of crime $\nabla_{f_w^1}\overline{C}$ for all directions $f_w^1 \in \mathcal{P}_0$.

For any wage $w \in \mathsf{supp}\, f_w^0$, let $\widehat{\mathbb{E}}[m|w, f_w^0]$ be the mean report of crime from monitors with wage w under policy f_w^0. Recall that the support of f_w^0 contains the support of all $f_w^1 \in \mathcal{P}_0$. Therefore, for any $f_w^1 \in \mathcal{P}_0$ we can construct synthetic mean reports of crime, under wage distribution f_w^1, *keeping the agents' bribing behavior constant* (i.e. optimal bribing given distribution of wages f_w^0), as follows:

$$R_0(f_w^1) \equiv \mathbb{E}_{f_w^0}\left[\widehat{\mathbb{E}}[m|w, f_w^0] \times \frac{f_w^1(w)}{f_w^0(w)}\right]. \tag{4}$$

Note that for all $f_w^1 \in \mathcal{P}_0$, synthetic mean reports $R_0(f_w^1)$ are computed using only reporting data generated at policy f_w^0. Indeed, $R_0(f_w^1)$ is obtained by simply re-weighting the original

reports $\widehat{\mathbb{E}}[m|w, f_w^0]$. The following result holds.

Proposition 4 (prior-free policy evaluation)**.** *There exists a fixed coefficient $\rho > 0$ such that for all $f_w^1 \in \mathcal{P}_0$,*

$$\nabla_{f_w^1} \overline{C} = \rho \left[\overline{R}_0 - R_0(f_w^1) \right].$$

This implies that a small movement from f_w^0 to f_w^1 will decrease crime ($\nabla_{f_w^1} \overline{C} < 0$) if and only if at policy f_w^0, the re-weighted reports of crime using distribution f_w^1 are larger than the original reports of crime.[16] In other words, it is optimal to move towards the policy f_w^1 such that, everything else equal, would maximize the amount of reported crime. The proof is instructive.

Proof. Take as given an arbitrary policy $f_w^1 \in \mathcal{P}_0$. Under wage schedule f_w^ϵ, the agent's payoff $U_A^\epsilon(\pi_A)$ 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_0 be the highest solution to this maximization problem for $\epsilon = 0$. Let π_A^0 denote the threshold at which agents are indifferent between engaging in crime or not under policy f_w^0. By the Envelope Theorem, $\forall \pi_A$,

$$\begin{aligned} \frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}\Big|_{\epsilon=0} &= \lambda(k - \tau_0) \left[\text{prob}_{f_w^1}(qw + \eta < \tau_0) - \text{prob}_{f_w^0}(qw + \eta < \tau_0) \right] \\ &= \lambda(k - \tau_0) \frac{1}{1 - F_{\pi_A}(\pi_A^0)} \left[\overline{R}_0 - R_0(f_w^1) \right], \end{aligned} \tag{5}$$

The second equality above follows from three observations. First, mean reports of crime $\overline{R}_0$ are equal to the product of baseline crime rates times the probability that equilibrium bribes

[16] These results relate the paper to a growing applied theory literature which studies mechanism design from the perspective of a principal with limited probabilistic sophistication. Responses to this challenges include solving for maxmin optimal designs (Hurwicz and Shapiro, 1978, Hartline and Roughgarden, 2008, Chassang, 2013, Frankel, 2014, Madarász and Prat, 2014, Prat, 2014, Carroll, 2013), as well as exploiting available data to discipline beliefs and guide policy design (Segal, 2003, Chassang and Padró i Miquel, 2013, Brooks, 2014).

are refused:

$$\overline{R}_0 = [1 - F_{\pi_A}(\pi_A^0)] \times [1 - \text{prob}_{f_w^0}(qw + \eta < \tau_0)].$$

Second, for any $\tilde{w} \in \mathsf{supp}\, f_w^0$, mean reports $\widehat{\mathbb{E}}[m|\tilde{w}, f_w^0]$ are equal to the product of baseline crime rates times the probability that a monitor with wage $\tilde{w}$ refuses the equilibrium bribe:

$$\begin{aligned} \forall \tilde{w} \in \mathsf{supp}\, f_w^0, \quad \widehat{\mathbb{E}}[m|\tilde{w}, f_w^0] &= [1 - F_{\pi_A}(\pi_A^0)] \times [1 - \text{prob}(q\tilde{w} + \eta < \tau_0)] \\ \Rightarrow \quad R_0(f_w^1) &= [1 - F_{\pi_A}(\pi_A^0)] \times [1 - \text{prob}_{f_w^1}(qw + \eta < \tau_0)]. \end{aligned}$$

Finally, for all $\epsilon \in [0, 1]$ let π_A^ϵ be the cutoff such that, under policy f_w^ϵ, an agent is criminal if and only if $\pi_A > \pi_A^\epsilon$. Note that

$$\begin{aligned} U_A^\epsilon(\pi_A^\epsilon) = \pi_A^\epsilon - k + \lambda \max_\tau (k - \tau) \text{prob}_{f_w^\epsilon}(qw + \eta < \tau) = 0 \\ \Rightarrow \quad \frac{\partial \pi_A^\epsilon}{\partial \epsilon} = -\frac{\partial}{\partial \epsilon} \lambda \max_\tau (k - \tau) \text{prob}_{f_w^\epsilon}(qw + \eta < \tau) = -\frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}. \end{aligned}$$

Since $\widehat{\mathbb{E}}[c^\epsilon | f_w^\epsilon] = 1 - F_{\pi_A}(\pi_A^\epsilon)$, it follows that $\nabla_{f_w^1} \overline{C} = f_{\pi_A}(\pi_A^0) \times \frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}\Big|_{\epsilon=0}$.

Combining these three observations, equation (5) implies that

$$\nabla_{f_w^1} \overline{C} = \frac{f_{\pi_A}(\pi_A^0)}{1 - F_{\pi_A}(\pi_A^0)} \lambda (k - \tau_0) \left[\overline{R}_0 - R_0(f_w^1)\right]$$

which proves Proposition 4. ■

Whenever distribution f_w^0 has full-support over $[\underline{w}, \overline{w}]$, this result lets us identify the optimal direction f_w^1 in which to move policy among all distributions with the same support. The proof also clarifies that even if f_w^0 does not have full support, one can form report measures $R_0(f_w^1)$ using data from an inexpensive experiment randomizing the wages of a small subset of monitors. One need not wait for equilibrium bribes and crime to adjust in order to interpret the data obtained from such an experiment. Indeed, the reason we can

evaluate local policy changes is precisely because the equilibrium response of criminal agents has a second order effect on their payoffs. Partial equilibrium responses provide the same data.

The fact that identification does not rely on costly experiments suggests the following process of continuous policy improvement. Starting from a policy f_w^0, one can engage in gradient-descent by iteratively picking directions for policy improvement $(\hat{f}_w^k)_{k\in\mathbb{N}}$ recursively defined by

$$\begin{aligned} \hat{f}_w^0 &\in \arg\min_{f_w\in\mathcal{P}_0} \overline{R}_0 - R_0(f_w) \\ \forall k \geq 1, \quad \hat{f}_w^k &\in \arg\min_{f_w\in\mathcal{P}_k} \overline{R}_k - R_k(f_w), \end{aligned}$$

where, for each $k \geq 1$, $\overline{R}_k$ is the proportion of monitors reporting crime under policy $f_w^k = (1-\epsilon)f_w^{k-1} + \epsilon\hat{f}_w^{k-1}$, and $R_k(f_w)$ are synthetic reports calculated as in (4) (with distribution f_w^k in place of f_w^0). Once the gradient is null in every direction, we have reached a local policy optimum.

5 Discussion

We explored the idea that random incentives can limit the cost of corruption by making side-contracting between criminal agents and monitors more difficult. We show that while the optimality of random incentives depends on unobserved pre-existing patterns of private information, it is possible to use naturally occuring data to guide policy choice. We now briefly discuss several extensions whose full treatment is relegated to Appendix A, and delineate what we think are the steps needed for a credible empirical evaluation of our policy recommendations.

5.1 Extensions

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

Arbitrary contracting between the principal and the monitor. Throughout the paper we assumed that the monitor is compensated with a fixed wage contract w and gets fired if she is caught misreporting. Under this assumption, Section 2 shows that deterministic incentive schemes are expensive under collusion, and that the principal can significantly reduce the cost of deterring crime by randomizing the monitor's wage. These results continue to hold if the principal can use arbitrary contracts to compensate the monitor. With more sophisticated contracts, the principal can reduce the cost of deterring crime by offering the monitor a higher compensation whenever she sends report $m = 1$. Indeed, a high compensation following report $m = 1$ increases the agent's cost of bribing the monitor, and remains cheap for the principal because it tends to be paid off of the equilibrium path. However, our assumption that reports are only partially verifiable (i.e. false reports are only detected with probability q) limits the extent to which the principal can exploit such incentives. With partially verifiable reports, as the monitor's compensation following message $m = 1$ gets large, it becomes optimal for her to report crime regardless of the agent's action. As a result, the cost of deterring crime with deterministic incentives remains high, and, as we show in Appendix A, the cost of keeping the agent non-criminal can be significantly reduced by using random incentives.

Multiple monitors. Section 2 shows that deterministic incentive schemes are undermined by the possibility of collusion. 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 an example with no 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, endogenous asymmetric information remains an effective way to reduce monitoring costs.

Extortion. The models of Sections 2 and 3 assume that the monitor sends a subgame-perfect message following disagreement at the side-contracting stage. This implies that the monitor can never extract bribes from an agent which she observes to be non-criminal. As Olken and Pande (2012) highlight, this prediction is frequently violated: non-criminal agents often have to pay bribes. A simple variation of our baseline model accounts for this. Assume that when the monitor has the bargaining power, she is able to commit to the message she would send in the event of a bargaining failure. A monitor can then extract rents from an non-criminal agent by committing to report the agent as criminal unless a bribe is paid. While this changes the agent's incentives to engage in crime, we show in Appendix A that our main results continue to hold in this setting: random incentives may reduce the cost of corruption, and it is possible to perform local policy evaluation on the basis of conditional report data.

Dynamic incentives. The model of Sections 2 and 3 is static. In practice, 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 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 policy evaluation results of Proposition 4 would extend. Remarkably, we are able to show in Appendix A that whenever the monitor's type η is persistent, Proposition 4 extends as is.

Participation constraints. Throughout the paper we assume that the monitor is risk-neutral, so that randomness in wages does not make participation constraints more difficult to satisfy. Risk-aversion on the monitor's side may restrain the use of random wages, but our qualitative results continue to hold in that case. The reason for this is that under collusion, participation is not binding. Indeed, in Section 2 we show that the cost of keeping the agent non-criminal with deterministic incentives is equal to $\frac{\pi_A}{q}$, compared to an outside option of 0. This means that the principal can use random incentives without affecting the monitor's participation constraint.

5.2 Steps towards 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 practical issue, we believe 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.

Heterogenous incentives without random wages. While randomizing wages is conceptually very simple, it does present significant practical challenges. In particular, it has distributional implications which stakeholders may find very unfair. We propose two ways to alleviate this concern while still generating appropriate heterogeneity in incentives.

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 adverse distributional consequences for non-corrupt monitors. 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.

Alternatively, one may be able to generate heterogenous incentives without randomization by letting the monitor's wage depend deterministically on data that is observable to the principal and the monitor, but not the agent. For instance, wages may be contingent on the monitor's tenure, diplomas, the number of crimes she has reported in the past, and so on... Such compensation schemes also introduce heterogeneity in the monitors' incentives, making side-contracting more difficult than under schemes that reward monitors with constant wages.

Picking a candidate policy. 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, Proposition 4 provides guidance on what alternative policy to choose using report data from any random incentive trial with a sufficiently rich support: choose the distribution that maximizes reports of crime keeping average wages constant. This implies that one can form a plausible candidate policy using report data from a single pilot intervention using any arbitrary full support distribution of wages.

Appendix

A Extensions

We now present several extensions describing how our results continue to hold in settings allowing for: more sophisticated contracting, multiple monitors, arbitrary bargaining mechanisms, extortion from non-criminal agents, repeated interaction, and changes in the timing of decisions.

A.1 Efficient contracting between the principal and monitor

Throughout the paper we assume that the principal compensates the monitor with an efficiency wage contract. This appendix shows how our results extend when we allow for arbitrary contracts. We consider the same environment as in Section 2, with one minor modification: we impose a participation constraint that the agent's payoff cannot be negative. We stress, however, that the results in the main text would remain unchanged if we added this constraint.[17] We also assume that the cost k_0 that a non-criminal agent expects from the judiciary is strictly positive.

Let $s \in \{\emptyset, f\}$ denote the signal that the principal observes by scrutinizing the monitor's report: the principal observes signal $s = f$ when she detects that the monitor's report is false, and observes signal $s = \emptyset$ otherwise.[18] The principal offers a wage contract $w(m, s)$ to the monitor, which determines the monitor's compensation as a function of the report she sends and the principal's signal. By limited liability, $w(m, s) \geq 0$ for all $(m, s) \in \{0, 1\} \times \{\emptyset, f\}$.

We begin with some preliminary results.

Lemma A.1. *Suppose the monitor is compensated with contract $w(m, s)$. Then, the monitor accepts a bribe τ from a criminal agent if and only if $\tau > w(1, \emptyset) - (1 - q)w(0, \emptyset) - qw(0, f)$.*

Proof. The monitor's payoff from accepting a bribe τ from a criminal agent is $\tau + (1 - q)w(0, \emptyset) + qw(0, f)$, while her payoff from rejecting the offer and sending a truthful message is $w(1, \emptyset)$. The agent accepts bribe τ if and only if $\tau > w(1, \emptyset) - (1-q)w(0, \emptyset) - qw(0, f)$. ■

Lemma A.2. *Let $w(m, s)$ be a contract that induces the monitor to send message $m = 0$ when the agent takes action $c = 0$ and offers bribe $\tau = 0$. Then, it must be that $w(0, \emptyset) \geq (1 - q)w(1, \emptyset) + qw(1, f)$.*

[17] Indeed, when the monitor is compensated with an efficiency wage $w \geq 0$ the agent can guarantee herself a payoff of 0 by taking action $c = 0$. When we allow for arbitrary contracts, the agent's participation constraint rules out wage structures under which the agent needs to bribe the monitor to get a favorable report after taking action $c = 0$.

[18] When the monitor sends report $m \neq c$, the principal observes signal $s = f$ with probability q and signal $s = \emptyset$ with probability $1 - q$. When the monitor sends report $m = c$, the principal observes signal $s = \emptyset$ with probability 1.

Proof. When the agent takes action $c = 0$ and offers bribe $\tau = 0$, the monitor's payoff from sending message $m = 0$ is $w(0, \emptyset)$, while her payoff from sending message $m = 1$ is $(1 - q)w(1, \emptyset) + qw(1, f)$. The monitor sends message $m = c = 0$ if and only if $w(0, \emptyset) \geq (1 - q)w(1, \emptyset) + qw(1, f)$. ■

Lemma A.3. *Under an optimal incentive scheme (either deterministic or random), a principal who wants to induce the agent to take action $c = 0$ offers the monitor contracts $w(m, s)$ with $w(0, \emptyset) = (1 - q)w(1, \emptyset)$ and $w(m, f) = 0$ for $m = 0, 1$.*

Proof. Suppose the incentive scheme induces the agent to take action $c = 0$ and satisfies the agent's participation constraint. By Lemma A.2, any contract $w(m, s)$ that the principal offers to the monitor with positive probability must satisfy $w(0, \emptyset) \geq (1-q)w(1, \emptyset)+qw(1, f)$; otherwise the agent's expected payoff from action $c = 0$ would be strictly negative, either because with positive probability the monitor sends a false report $m = 1$, or because the agent needs to bribe the monitor for a report $m = 0$. In either case, this would violate the agent's participation constraint.

This implies that under an optimal incentive scheme that induces the agent to take action $c = 0$, on the equilibrium path the monitor sends report $m = 0$ and receives a wage $w(0, \emptyset)$. If $w(0, \emptyset) > (1 - q)w(1, \emptyset) + qw(1, f)$ for some contract $w(m, s)$ that is offered with positive probability, the principal would be strictly better-off by reducing $w(0, \emptyset)$ as this would reduce wage payments and would also increase the cost of bribing the monitor (Lemma A.1).

By limited liability it must be that $w(m, f) \geq 0$ for $m = 0, 1$. Setting $w(0, f) = 0$ is optimal as it increases the cost of bribing the monitor. Finally, since $w(0, \emptyset) = (1 - q)w(1, \emptyset) + qw(1, f)$, setting $w(1, f) = 0$ reduces the wage $w(0, \emptyset)$ that the principal pays on the equilibrium path and also increases the cost of bribing the monitor. ■

We now consider the case in which the principal compensates the agent with a deterministic contract $w(m, s)$. The following result generalizes Lemma 2 to the current setting.

Lemma A.4. *Suppose the principal uses a deterministic contract $w(m,s)$. Under collusion, the minimum cost of wages needed to induce the agent to be non-criminal is equal to $\frac{1-q}{2-q}\frac{\pi_A}{q}$.*

Proof. By Lemmas A.1 and A.3, the monitor accepts a bribe τ from a criminal agent if and only if $\tau > w(1,\emptyset) - (1-q)w(0,\emptyset)$. The agent's payoff from taking action $c=1$ is then $\pi_A - \min\{k, w(1,\emptyset) - (1-q)w(0,\emptyset)\}$, while her payoff from taking action $c=0$ is 0. To induce the agent to take action $c=0$, it must be that $w(1,\emptyset) - (1-q)w(0,\emptyset) \geq \pi_A$. By Lemma A.3, $w(0,\emptyset) = (1-q)w(1,\emptyset)$, so the previous inequality yields $w(0,\emptyset) \geq \frac{1-q}{2-q}\frac{\pi_A}{q}$. ■

Consider next the case in which the principal randomizes over the monitor's contract $w(m,s)$. By Lemma A.3, it is optimal for the principal to offer contracts $w(m,s)$ such that $w(0,\emptyset) = (1-q)w(1,\emptyset)$ and $w(m,f) = 0$ for $m = 0,1$. Therefore, it is without loss of optimality to focus on distributions over wages $w(0,\emptyset)$, with the understanding that a contract with $w(0,\emptyset) = w \geq 0$ has $w(1,\emptyset) = \frac{w}{1-q}$ and $w(m,f) = 0$ for $m = 0,1$.

The following result generalizes Proposition 1 to the current setting.

Proposition A.1. *Under collusion, it is optimal for the principal to use random contracts. The cost-minimizing distribution $\hat{F}_w^*$ over wages $w(0,\emptyset)$ that induces the agent to be non-criminal is described by*

$$\forall w \in \left[0, \frac{\pi_A}{q}\frac{1-q}{2-q}\right], \quad \hat{F}_w^*(w) = \frac{k-\pi_A}{k - qw\frac{2-q}{1-q}}. \tag{6}$$

The corresponding cost of wages $\hat{W}^*(\pi_A) \equiv \mathbb{E}_{\hat{F}^*}[w]$ is

$$\hat{W}^*(\pi_A) = \frac{1-q}{2-q}\frac{\pi_A}{q}\left[1 - \frac{k-\pi_A}{\pi_A}\log\left(1 + \frac{\pi_A}{k-\pi_A}\right)\right] < \frac{1-q}{2-q}\frac{\pi_A}{q}\frac{\pi_A}{k}. \tag{7}$$

Proof. By Lemma A.1, a monitor with contract $w(m,s)$ accepts a bribe τ from a criminal agent if and only if $\tau > w(1,\emptyset) - (1-q)w(0,\emptyset) - qw(0,f) = \frac{2-q}{1-q}qw(0,\emptyset)$, where the last equality follows since $w(1,\emptyset) = \frac{w(0,\emptyset)}{1-q}$ and $w(m,f) = 0$ for $m = 0,1$ (Lemma A.3). A distribution F over wages $w(0,\emptyset)$ induces the agent to take action $c=0$ if and only if, for

every bribe offer $\tau \geq 0$, $\pi_A - k + (k - \tau)\text{prob}(\tau > \frac{2-q}{1-q} qw) \leq 0$, or equivalently, if and only if, for every $\tau \geq 0$, $F\left(\frac{\tau}{q}\frac{1-q}{2-q}\right) \leq \frac{k-\pi_A}{k-\tau}$. Using the change in variable $w = \frac{\tau}{q}\frac{1-q}{2-q}$, we obtain that wage distribution F induces the agent to take action $c = 0$ if and only if,

$$\forall w \in \left[0, \frac{\pi_A}{q}\frac{1-q}{2-q}\right], \quad F(w) \leq \frac{k - \pi_A}{k - qw\frac{2-q}{1-q}}. \tag{8}$$

By first-order stochastic dominance, it follows that in order to minimize expected wages, the optimal distribution must satisfy (8) with equality. This implies that the optimal wage distribution is described by (6). Expected cost expression (7) follows from integration and straightforward computations. ■

A.2 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 chooses whether or not to engage in crime $c \in \{0, 1\}$, 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 criminal agents and (for simplicity) a cost of 0 on non-criminal 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 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 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 chooses an action $c \in \{0, 1\}$;
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 Lemma 2 to the current setting.

Lemma A.5. *Assume that the principal hires two monitors and uses deterministic wages. Under* no collusion *the principal can induce the agent to be non-criminal at 0 cost.*

Under collusion*, the minimum cost of wages needed to induce the agent to be non-criminal 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 of a criminal agent is $\pi_A - k < 0$, while her payoff when non-criminal is 0.[19]

Consider next the case of *collusion*. Solving the game by backward induction, if a criminal 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 criminal agent who bribes both monitors is $\pi_A - 2qw$. The payoff of a non-criminal agent is 0, so the agent will be non-criminal 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-criminal is $\frac{\pi_A}{q}$. ■

[19]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.3 Arbitrary bargaining

The model of Sections 2 and 3 simplifies the side-contracting stage by assuming take-it-or-leave-it offers. This appendix allows for arbitrary bargaining mechanisms. We study a model 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 crime is $U_A = \pi_A - k + \int (P(w)k - \tau(w))\, dF(w)$. The individual rationality constraint of a criminal agent is $U_A \geq \pi_A - k$, since a criminal 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 criminal 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 agent is criminal is

$$(1 - \lambda) \int U_M(w) dF(w) + \lambda U_A, \tag{9}$$

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 (9). 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 criminal agent gets under a bargaining mechanism in $\Gamma(F,\lambda)$. The agent has an incentive to be non-criminal if $\tilde{U}_A(F,\lambda) \leq 0$.

The following result generalizes Proposition 1 to this setting.

Proposition A.2. *Suppose that, at the collusion stage, the monitor and the agent use an incentive compatible and individually rational mechanism that maximizes* (9).

(i) If $\lambda \in (1/2,1]$, the cost minimizing wage distribution $\tilde{F}_w^*$ that induces the agent to be non-criminal is described by

$$\forall w \in [0, \pi_A/q], \quad \tilde{F}_w^*(w) = \left(\frac{k-\pi_A}{k-qw}\right)^{\frac{2\lambda-1}{\lambda}}. \tag{10}$$

(ii) If $\lambda \in [0,1/2]$, the cost minimizing wage distribution $\tilde{F}_w^*$ that induces the agent to be non-criminal has $\tilde{F}_w^*(0) = 1$.

Proof. By standard arguments, any incentive compatible mechanism (P,τ) must satisfy: (i) $P(w)$ is decreasing, and (ii) $U_M'(w) = 1 - qP(w)$ a.e.. 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 criminal is

$$
\begin{aligned}
&(1-\lambda)\int_{\underline{w}}^{\overline{w}} U_M(w)dF(w) + \lambda U_A \\
&= \int_{\underline{w}}^{\overline{w}} \left[(1-\lambda)(\tau(w) + (1-qP(w))w) + \lambda(P(w)k - \tau(w))\right] dF(w) + \lambda(\pi_A - k) \\
&= \int_{\underline{w}}^{\overline{w}} \left[P(w)\lambda\left(k - qw\right) + (1-\lambda)w\right] dF(w) + \lambda(\pi_A - k) + (1-2\lambda)\left(\int_{\underline{w}}^{\overline{w}} qP(w)F(w)dw + c\right).
\end{aligned} \tag{11}
$$

We use the following lemma.

Lemma A.6. *For all $\lambda \in (1/2, 1]$, the mechanism (P, τ) that maximizes* (11) *has: (i) $P(w) = 1$ if $w < w^*$ and $P(w) = 0$ if $w > w^*$ for some $w^* \in [\underline{w}, \overline{w}]$, and (ii) $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}$.*

Proof. Note first that (11) is maximized by setting $c = 0$ when $\lambda \in (1/2, 1]$. Moreover, when $\lambda \in (1/2, 1]$ any mechanism (P, τ) that maximizes (11) must be such $P(w) = 0$ for all $w \geq k/q$.

We now show that the mechanism that maximize (11) is such that $P(w)$ only takes values 0 or 1. From above, we know that $P(w) = 0$ for all $w \geq k/q$. Suppose by contradiction that there exists an interval $V \subset [0, k/q]$ 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 (11) larger. If $H < 0$, decreasing $P(w)$ over this interval (subject to the constraint that P is decreasing) also makes (11) larger. Such improvements are exhausted when $P(w)$ only takes values 0 and 1.[20] 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, since (11) is maximized by setting

[20]Note that these changes in $P(w)$ do not conflict with the participation constraints of monitor and agent. Indeed, $U_M(w) = \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} + w \geq w$ for any incentive compatible mechanism (P, τ). Moreover, for all w, $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} \leq P(w)k$, where the inequality follows since any mechanism that maximizes (11) has $P(w) = 0$ for all $w \geq k/q$ and since $P(\cdot)$ is decreasing. Hence, $U_A = \pi_A - k + \int (P(w)k - \tau(w))dF(w) \geq \pi_A - k$.

$c = 0$ when $\lambda \in (1/2, 1]$, $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}$. Since $P(w) = 1$ if $w < w^*$ and $P(w) = 0$ if $w > w^*$, it follows that $\tau(w) = qw^*$ if $w < w^*$ and $\tau(w) = 0$ if $w > w^*$. ■

We now conclude the proof of Proposition A.2, beginning with point (i). Fix $\lambda \in (1/2, 1]$ and let F be a cost-minimizing wage schedule that induces the agent to be non-criminal. Let (P, τ) be the mechanism that maximizes the weighted sum of payoffs (11) under distribution F. By Lemma A.6, $P(w) = \mathbf{1}_{w \leq w^*}$ and $\tau(w) = qw^*\mathbf{1}_{w \leq w^*}$ for some w^*. Under this mechanism (11) becomes

$$\begin{aligned}
&\lambda \left[F(w^*)k - \int_0^{w^*} qwdF(w) + \pi_A - k \right] + (1-\lambda) \int wdF(w) + (1-2\lambda) \int_0^{w^*} qF(w)dw \\
=&\lambda \left[F(w^*)(k - qw^*) + \pi_A - k \right] + (1-\lambda) \int wdF(w) + (1-\lambda) \int_0^{w^*} qF(w)dw,
\end{aligned}$$

where we used $\int_0^{w^*} qwdF(w) = qw^*F(w^*) - \int_0^{w^*} qF(w)dw$. Since (P, τ) maximizes the weighted sum of payoffs, for all $\hat{w} \neq w^*$ it must be that

$$\lambda F(w^*)(k - qw^*) + (1-\lambda) \int_0^{w^*} qF(w)dw \geq \lambda F(\hat{w})(k - q\hat{w}) + (1-\lambda) \int_0^{\hat{w}} qF(w)dw$$

Otherwise, if the inequality 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 $w > \hat{w}$.

For any $\hat{w} \in \mathsf{supp}\, F$, let $(P_{\hat{w}}, \tau_{\hat{w}})$ be the mechanism with $P_{\hat{w}}(w) = \mathbf{1}_{\{w \leq \hat{w}\}}$ and $\tau_{\hat{w}}(w) = \mathbf{1}_{\{w \leq \hat{w}\}} q\hat{w}$. Recall that $\Gamma(F, \lambda)$ is the set of bargaining mechanisms that maximize (11) and that $\tilde{U}_A(F, \lambda)$ is the lowest utility that a criminal agent gets under a mechanism in $\Gamma(F, \lambda)$. By our arguments above,

$$\Gamma(F, \lambda) = \left\{ (P_{\hat{w}}, \tau_{\hat{w}}) : \hat{w} \in \arg\max_{w'} \lambda F(w')(k - qw') + (1-\lambda) \int_0^{w'} qF(w)dw \right\}.$$

Suppose that there exists w_1 and $w_2 > w_1$ such that $(P_w, \tau_w) \in \Gamma(F, \lambda)$ for $w = w_1, w_2$. Note that the agent's payoff from being criminal under mechanism (P_w, τ_w) is $F(w)(k - qw) + \pi_A - k$.

Since $(P_w, \tau_w) \in \Gamma(F, \lambda)$ for $w = w_1, w_2$,

$$\lambda F(w_1)(k - qw_1) + (1-\lambda)\int_0^{w_1} qF(w)dw = \lambda F(w_2)(k - qw_2) + (1-\lambda)\int_0^{w_2} qF(w)dw$$

and so $F(w_2)(k-qw_2) < F(w_1)(k-qw_1)$. This implies that, $\tilde{U}_A(F,\lambda) = F(\tilde{w})(k-q\tilde{w})+\pi_A-k$, where $\tilde{w} \equiv \sup\{\hat{w} \in \mathsf{supp}\, F : \hat{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw\}$. Since F induces the agent to be non-criminal, $\tilde{U}_A(F,\lambda) = F(\tilde{w})(k-q\tilde{w}) + \pi_A - k \leq 0$.

Let $\overline{w}$ be the highest wage in the support of F. We now show that, if F is an optimal distribution, it must be that $\overline{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw$. Suppose by contradiction that this is not true, so that $\overline{w} > \tilde{w} = \sup\{\hat{w} \in \mathsf{supp}\, F : \hat{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw\}$. Pick $\epsilon \in (0, \overline{w} - \tilde{w})$ small and let F^ϵ be a c.d.f. with $F^\epsilon(w) = F(w)$ for all $w < \overline{w} - \epsilon$ and $F^\epsilon(\overline{w}-\epsilon) = 1$. By first-order stochastic dominance, $\mathbb{E}_{F^\epsilon}[w] < \mathbb{E}_F[w]$. By the definition of $\tilde{w}$,

$$\lambda F(\tilde{w})(k-q\tilde{w}) + (1-\lambda)\int_0^{\tilde{w}} qF(w)dw \geq \lambda F(\hat{w})(k-q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF(w)dw,$$

for all $\hat{w}$, with strict inequality for all $\hat{w} \in (\tilde{w}, \overline{w}]$. Therefore, there exists $\epsilon > 0$ small enough such that, for all $\hat{w}$,

$$\lambda F^\epsilon(\tilde{w})(k-q\tilde{w}) + (1-\lambda)\int_0^{\tilde{w}} qF^\epsilon(w)dw \geq \lambda F^\epsilon(\hat{w})(k-q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF^\epsilon(w)dw$$

This implies that mechanism $(P_{\tilde{w}}, \tau_{\tilde{w}})$ is still optimal under distribution F^ϵ, and so $\tilde{U}_A(F^\epsilon, \lambda) \leq F(\tilde{w})(k-q\tilde{w}) + \pi_A - k \leq 0$. But this cannot be, since F is a cost-minimizing distribution that induces the agent to be non-criminal. Therefore, if F is optimal it must be that $\overline{w} = \sup\{\hat{w} \in \mathsf{supp}\, F : \hat{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw\}$. The agent's payoff from being criminal under mechanism $(P_{\overline{w}}, \tau_{\overline{w}})$ is $k - q\overline{w} + \pi_A - k \leq 0 \iff \overline{w} \geq \frac{\pi_A}{q}$.

By the arguments above, for all $\hat{w} \in [0, \overline{w}]$,

$$\lambda(k - q\overline{w}) + (1-\lambda)\int_0^{\overline{w}} qF(w)dw \geq \lambda F(\hat{w})(k - q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF(w)dw$$

$$\iff \lambda(k - q\overline{w}) + (1-\lambda)\int_{\hat{w}}^{\overline{w}} qF(w)dw \geq \lambda F(\hat{w})(k - q\hat{w}) \tag{12}$$

We now show that, if F is an optimal distribution, (12) must hold with equality for all $\hat{w} \in [0, \overline{w}]$. Suppose by contradiction that there is an interval $[w_1, w_2] \subset [0, \overline{w})$ such that (12) is slack for all $\hat{w} \in [w_1, w_2]$. By first-order stochastic dominance, increasing $F(\cdot)$ over $[w_1, w_2]$ (subject to the constraint that F is increasing) reduces expected wage payments. Moreover, increasing $F(\cdot)$ over $[w_1, w_2]$ relaxes (12) for all $\hat{w} < w_1$ and does not affect (12) for all $\hat{w} > w_2$. This implies that mechanism $(P_{\overline{w}}, \tau_{\overline{w}})$ still maximizes the weighted sum of payoffs (11) after increasing $F(\cdot)$ slightly over $[w_1, w_2]$, and so the agent's payoff from being criminal is $k - q\overline{w} + \pi_A - k \leq 0$. But this cannot be, since F is a cost-minimizing distribution that induces the agent to be non-criminal. Therefore, if F is optimal, (12) must hold with equality for all $\hat{w} \leq \overline{w}$.

Since (12) holds with equality for all $\hat{w} \leq \overline{w}$, $\lambda F(\hat{w})(k - q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF(w)dw$ is constant over $[0, \overline{w}]$. Differentiating this expression with respect to $\hat{w}$, it must be that

$$F'(\hat{w})\lambda[k - q\hat{w}] + qF(\hat{w})(1 - 2\lambda) = 0. \tag{13}$$

The solution to the differential equation (13) is $F(w) = C\left(\frac{1}{k-qw}\right)^{\frac{2\lambda-1}{\lambda}}$, where C is a constant such that $F(\overline{w}) = 1$; i.e., $C = (k - q\overline{w})^{\frac{2\lambda-1}{\lambda}}$. Finally, by our arguments above, under distribution F the agent will have an incentive to be non-criminal as long as $k - q\overline{w} + \pi_A - k \leq 0 \iff \overline{w} \geq \frac{\pi_A}{q}$. Since the constant C is decreasing in $\overline{w}$, an optimal distribution must have $\overline{w} = \frac{\pi_A}{q}$. Hence, $C = (k - \pi_A)^{\frac{2\lambda-1}{\lambda}}$, so the optimal distribution is (10).

We now turn to point (ii). When $\lambda \leq 1/2$, the mechanism (P, τ) that maximizes (11) must make the constant c as large as possible, subject to 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 engaging in crime under a mechanism that maximizes (11) is $\pi_A - k < 0$, regardless of the wage schedule. This implies that the agent has an incentive to be non-criminal even when F has all its mass at $w = 0$. ■

A.4 Extortion

This section shows how our results extend to settings in which the monitor can extort transfers from non-criminal agents 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 criminal agents and an expected cost $k_0 \in (0, k]$ on non-criminal agents.

Lemma A.7. *If the monitor acts as proposer when the agent is non-criminal, she demands a bribe $\tau = k_0$ if her type is $\eta < k_0$, and she demands no bribe (i.e. $\tau = 0$) if her type is $\eta \geq k_0$. A non-criminal agent accepts any offer $\tau \leq k_0$.*

Proof. Suppose the monitor makes an offer τ to a non-criminal agent and commits to sending a false message if her proposal is rejected. In this case, it is optimal for a non-criminal 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 $\eta < k_0$. ■

Lemma A.8. *If the monitor acts as a proposer at the collusion stage, she demands a bribe $\tau \geq k$ when the agent is criminal. A criminal agent accepts any offer $\tau \leq k$.*

Proof. The proof of Lemma A.8 is identical to the proof of Lemma 3. ■

Lemma A.7 implies that the payoff of a non-criminal agent is $-(1-\lambda)k_0F_\eta(k_0)$, while Lemma A.8 implies that the payoff of a criminal 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 criminal agent incurs when the monitor sends report $m=1$ by $k-(1-\lambda)k_0F_\eta(k_0)$. Note that this term does not depend on the distribution of wages. Hence, all the results in Sections 3 and 4 continue to hold when the monitor can commit to sending a false message.

A.5 Dynamic incentives

The model of Sections 2 and 3 assumes that the principal provides incentives to monitors by taking away one-shot wages in the event that misreporting is detected. 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 it is still possible to 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 discrete 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 the same 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 of Section 3. The only difference is that a monitor who is found misreporting receives her current period wage w but gets fired and therefore losses her continuation value from employment.

Let $W(w, \eta)$ be the value of remaining employed for a monitor with wage w and type η. We 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 criminal agent is $(1-\delta)(\tau - \eta) - q\delta W(w, \eta)$, where $\delta < 1$ is the discount factor. This implies the following.

Lemma A.9. *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 to Lemma 3 in the current setting.

Lemma A.10. *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 criminal, and a bribe $\tau = 0$ when the agent is non-criminal.

The agent accepts any offer $\tau \leq k$ when she is criminal and any offer $\tau = 0$ when she is non-criminal.

Lemma A.10 implies that the payoff of a non-criminal agent is 0. Moreover, by Lemmas A.9 and A.10 the payoff of a criminal agent of type π_A is

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

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

Lemma A.11. *Let $\tau^* \leq k$ be the offer that criminal agents make. A monitor with type η and wage w accepts offer τ^* from a criminal 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 criminal agent. The value function of this monitor is

$$\begin{aligned} W(w,\eta) &= (1-\delta)w + \delta W(w,\eta) + (1-\lambda)\overline{C}\left((1-\delta)(k-\eta) - q\delta W(w,\eta)\right) \Rightarrow \\ W(w,\eta) &= \frac{(1-\delta)(w+(1-\lambda)(k-\eta)\overline{C})}{1-\delta+\delta q(1-\lambda)\overline{C}}. \end{aligned} \tag{14}$$

The last term in the first expression is the net payoff that the monitor gets when she is proposer against a criminal agent: with probability $(1-\lambda)\overline{C}$ the monitor is proposer against a criminal agent, extracts a bribe k in exchange of a false message, pays the idiosyncratic cost η and is fired with probability q.[21] Since this monitor is indifferent between accepting offer τ^* or rejecting it, $(1-\delta)(\tau^*-\eta) = q\delta W(w,\eta)$, which by equation (14) 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 4 extends to this setting. Take as given a distribution of wages with cdf F_w^0 and density f_w^0. For any alternative policy f_w^1, construct the mixture $f_w^\epsilon = (1-\epsilon)f_w^0 + \epsilon f_w^1$ and let $\overline{C}^\epsilon = \widehat{\mathbb{E}}[c^\epsilon | f_w^\epsilon]$ be the proportion of criminal agents under policy f_w^ϵ. Recall that $\nabla_{f_w^1}\overline{C}$ is the gradient of equilibrium crime in policy direction f_w^1:

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

As in Section 4, for any $w \in \mathsf{supp}\, f_w^0$, let $\widehat{\mathbb{E}}[m|w, f_w^0]$ be the mean report of corruption

[21] 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 criminal agent.

from monitors with wage w under policy f_w^0. Recall that $\mathcal{P}_0$ is the set of policies f_w^1 such that $\mathsf{supp}\, f_w^1 \subset \mathsf{supp}\, f_w^0$ and $\mathbb{E}_{f_w^1}[w] = \mathbb{E}_{f_w^0}[w]$. For any $f_w^1 \in \mathcal{P}_0$, let

$$R_0(f_w^1) \equiv \mathbb{E}_{f_w^0}\left[\widehat{\mathbb{E}}[m|w, f_w^0] \times \frac{f_w^1(w)}{f_w^0(w)}\right].$$

Proposition A.3. *There exists a fixed coefficient $\rho > 0$ such that, for all $f_w^1 \in \mathcal{P}_0$,*

$$\nabla_{f_w^1}\overline{C} = \rho\left[\overline{R}_0 - R_0(f_w^1)\right].$$

Proof. Let τ_ϵ be the optimal offer by a criminal agent under policy f_w^ϵ. Let π_A^0 denote the threshold at which agents are indifferent between engaging in crime or not under policy f_w^0.

By Lemma A.11, the probability that a monitor accepts offer τ_ϵ under policy f_w^ϵ is $\text{prob}_{f_w^\epsilon}\left(\eta < \overline{\eta}(\tau_\epsilon, w, \overline{C}^\epsilon)\right) = \mathbb{E}_{f_w^\epsilon}[F_\eta(\overline{\eta}(\tau_\epsilon, w, \overline{C}^\epsilon))]$. The payoff of a criminal agent with type π_A under policy f_w^ϵ is

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

By the Envelope Theorem,

$$\begin{aligned}\frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}\Big|_{\epsilon=0} \;=\;& \lambda(k-\tau_0)\left[\mathbb{E}_{f_w^1}\left[F_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0))\right] - \mathbb{E}_{f_w^0}\left[F_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0))\right]\right] \\ &+\lambda(k-\tau_0)\mathbb{E}_{f_w^0}\left[f_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0)) \times \frac{\partial\overline{\eta}(\tau_0, w, \overline{C}^0)}{\partial\overline{C}^0}\nabla_{f_w^1}\overline{C}\right].\end{aligned}$$

The equation above can be written as

$$\begin{aligned}&\frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}\Big|_{\epsilon=0} - \lambda(k-\tau_0)\mathbb{E}_{f_w^0}\left[f_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0)) \times \frac{\partial\overline{\eta}(\tau_0, w, \overline{C}^0)}{\partial\overline{C}^0}\nabla_{f_w^1}\overline{C}\right] \\ &= \lambda(k-\tau_0)\left[\mathbb{E}_{f_w^1}\left[F_\eta(\overline{\eta}(\tau, w, \overline{C}^0))\right] - \mathbb{E}_{f_w^0}\left[F_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0))\right]\right] \\ &= \frac{\lambda(k-\tau_0)}{1 - F_{\pi_A}(\pi_A^0)}[\overline{R}_0 - R_0(f_w^1)] \qquad (15)\end{aligned}$$

The last equality in equation (15) follows from two observations. First, mean reports of crime are equal to the product of baseline crime rates times the probability that equilibrium bribes are refused:

$$\overline{R}_0 = [1 - F_{\pi_A}(\pi_A^0)] \times \left[1 - \mathbb{E}_{f_w^0}\left[F_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0))\right]\right].$$

Second, for any $\tilde{w} \in \mathsf{supp}\, f_w^0$, mean reports $\widehat{\mathbb{E}}[m|\tilde{w}, f_w^0]$ are equal to the product of baseline crime rates times the probability that a monitor with wage $\tilde{w}$ refuses the equilibrium bribe:

$$\begin{aligned}\forall \tilde{w} \in \mathsf{supp}\, f_w^0, \quad \widehat{\mathbb{E}}[m|\tilde{w}, f_w^0] &= [1 - F_{\pi_A}(\pi_A^0)] \times [1 - F_\eta(\overline{\eta}(\tau_0, \tilde{w}, \overline{C}^0))] \\ &\Rightarrow R_0(f_w^1) = [1 - F_{\pi_A}(\pi_A^0)] \times \left[1 - \mathbb{E}_{f_w^1}\left[F_\eta(\overline{\eta}(\tau, w, \overline{C}^0))\right]\right].\end{aligned}$$

Finally, note that $\frac{\partial \overline{\eta}(\tau_0, w, \overline{C}^0)}{\partial \overline{C}^0} = \frac{(\tau_0 - k)\delta q(1-\lambda)}{1-\delta} < 0$. Since $\nabla_{f_w^1} \overline{C} = f_{\pi_A}(\pi_A^0) \times \frac{\partial U_A^\epsilon(\pi_A)}{\partial \epsilon}\Big|_{\epsilon=0}$, it follows that

$$\begin{aligned}&\nabla_{f_w^1} \overline{C} \times \left[1 + f_{\pi_A}(\pi_A^0) \frac{\lambda(k - \tau_0)^2 \delta q(1-\lambda)}{1-\delta} \mathbb{E}_{f_w^0}\left[f_\eta(\overline{\eta}(\tau_0, w, \overline{C}^0))\right]\right] \\ &= \frac{f_{\pi_A}(\pi_A^0)}{1 - F_{\pi_A}(\pi_A^0)} \lambda(k - \tau_0)[\overline{R}_0 - R_0(f_w^1)].\end{aligned}$$

This completes the proof. ■

A.6 Alternative timing of decisions

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

We consider a model in which the agent chooses action $c \in \{0, 1\}$ after side-contracting with the monitor, but which is otherwise the same as the model in Section 2. At the side-contracting stage the agent makes a take-it-or-leave-it offer $\tau \geq 0$ to the monitor. If

the monitor accepts the agent's offer, she commits to send report $m = 0$ to the principal regardless of the agent's action. Otherwise, if the monitor rejects the agent's offer, she sends the report $m \in \{0, 1\}$ that maximizes her expected payoff. As in Section 2, the principal detects false messages with probability q. The monitor is compensated with an efficiency wage $w \geq 0$, and losses this wage if the principal detects that the message was false.

Lemma A.12. *The agent takes action $c = 1$ if and only if the monitor accepts her bribe. A monitor with wage w accepts a bribe τ if and only if $\tau > qw$.*

Proof. If the monitor accepts the agent's bribe τ, the agent's payoffs from action $c = 1$ is $\pi_A - \tau$, while her payoff from action $c = 0$ is $-\tau$. If the monitor rejects the agent's bribe, the agent's payoff from $c = 1$ is $\pi_A - k < 0$ (since in this case the monitor will find it optimal to send message $m = 1$), while her payoff from action $c = 0$ is 0. Therefore, the agent takes action $c = 1$ if and only if the monitor accepts her bribe.

By the previous paragraph, the monitor's payoff from accepting bribe τ is $\tau + (1 - q)w$, while her payoff from rejecting the bribe and sending a truthful message is w. The monitor finds it optimal to accept bribe τ if and only if $\tau > qw$. ■

We now consider the case in which the principal compensates the agent with a deterministic wage w. The following result generalizes Lemma 2 to the current setting; its proof is identical to the proof of Lemma 2 and hence omitted.

Lemma A.13. *Suppose the principal uses a deterministic wage w. Under collusion, the minimum cost of wages needed to induce the agent to take action $c = 0$ is equal to $\frac{\pi_A}{q}$.*

Consider next the case in which the principal randomizes over the monitor's wage. Suppose the principal pays the monitor an efficiency wage drawn from the c.d.f. F. Note that the agent's payoff from making an offer $\tau \geq 0$ is $F(\tau/q)(\pi_A - \tau) + (1 - F(\tau/q)) \times 0$. Let τ_F^* be the solution to $\max_\tau F(\tau/q)(\pi_A - \tau)$. For any distribution F, the principal's payoff is

$$F\left(\frac{\tau_F^*}{q}\right)\pi_P - \mathbb{E}_F[w].$$

Under wage distribution F, the monitor accepts the agent's bribe when her wage is lower than τ_F^*/q. In this case, the agent takes action $c = 1$ and the principal incurs cost $\pi_P < 0$.

Proposition A.4. *Assume that the agent and monitor collude before the agent chooses $c \in \{0, 1\}$. Then, the optimal wage distribution $\widetilde{F}^*$ is described by,*

$$\forall w \in \left[0, \frac{\pi_A}{q}\left(1 - e^{\frac{q\pi_P}{\pi_A}}\right)\right], \quad \widetilde{F}_w^*(w) = \frac{e^{\frac{q\pi_P}{\pi_A}}\pi_A}{\pi_A - qw}. \tag{16}$$

When the principal pays the monitor a wage drawn from $\widetilde{F}_w^*$, the agent takes action $c = 1$ with probability $\widetilde{F}_w^*(0) \in (0, 1)$.

Proof. Consider first distributions F such that $F\left(\frac{\tau_F^*}{q}\right) = 0$. Note that $F\left(\frac{\tau_F^*}{q}\right) = 0$ implies that $0 \geq \max_\tau F(\tau/q)(\pi_A - \tau)$, and so $F(\tau/q) = 0$ for all $\tau < \pi_A$. Therefore, for distributions F such that $F\left(\frac{\tau_F^*}{q}\right) = 0$, the minimum cost of wages is achieved with a distribution that puts all its mass at $w = \pi_A/q$. The principal's payoff under this distribution is $-\pi_A/q$. Our arguments below show that such a distribution is never optimal.

Consider next distributions F such that $F\left(\frac{\tau_F^*}{q}\right) > 0$. Since $\tau_F^* \geq 0$ is the optimal offer, for all $\tau \geq 0$,

$$F\left(\frac{\tau_F^*}{q}\right)(\pi_A - \tau_F^*) \geq F\left(\frac{\tau}{q}\right)(\pi_A - \tau) \iff F\left(\frac{\tau}{q}\right) \leq F\left(\frac{\tau_F^*}{q}\right)\frac{\pi_A - \tau_F^*}{\pi_A - \tau}. \tag{17}$$

By first order stochastic dominance, an optimal wage distribution F with $F\left(\frac{\tau_F^*}{q}\right) > 0$ must be such that (17) holds with equality for all τ such that $F(\tau/q) < 1$.

Next, we show that the optimal distribution F with $F\left(\frac{\tau_F^*}{q}\right) > 0$ must be such that $\tau_F^* = 0$. Let F be such that $\tau_F^* > 0$, and let $\hat{F}$ be an alternative distribution described by: $\hat{F}(0) = F(\tau_F^*/q)$ and $\hat{F}(\tau/q) = \frac{\hat{F}(0)\pi_A}{\pi_A - \tau}$ for all $\tau \in [0, \pi_A(1 - \hat{F}(0))]$. By construction, bribe $\tau = 0$ maximizes $\hat{F}(\tau/q)(\pi_A - \tau)$. Since $\hat{F}(0) = F(\tau_F^*/q)$, the probability that the agent takes action $c = 1$ is the same under $\hat{F}$ than under F. Moreover, for all τ such that $\hat{F}(\tau/q) < 1$, $\hat{F}(\tau/q) = \frac{\hat{F}(0)\pi_A}{\pi_A - \tau} > F(\tau_F^*/q)\frac{\pi_A - \tau_F^*}{\pi_A - \tau} \geq F(\tau/q)$ (where the last inequality follows since offer τ_F^*

is optimal under policy F). This implies that $\mathbb{E}_F[w] > \mathbb{E}_{\hat{F}}[w]$, so the principal's payoff is larger under $\hat{F}$ than under F.

Using the change in variable $w = \tau/q$, the two paragraphs above imply that the optimal wage distribution F with $F\left(\frac{\tau_F^*}{q}\right) > 0$ is such that $\tau_F^* = 0$ and is described by

$$\forall w \in \left[0, \frac{\pi_A}{q}(1 - F(0))\right], \quad F(w) = \frac{F(0)\pi_A}{\pi_A - qw}.$$

The principal's expected payoff from using this wage distribution is

$$F(0)\pi_P - \mathbb{E}_F[w] = F(0)\pi_P - \frac{\pi_A}{q}(1 - F(0) + F(0)\ln F(0)).$$

This expression is strictly concave in $F(0)$, and converges to $-\frac{\pi_A}{q}$ as $F(0) \to 0$. Maximizing this expression with respect to $F(0)$ yields $F(0) = e^{\frac{q\pi_P}{\pi_A}} \in (0, 1)$. Therefore, the optimal wage distribution is given by (16). ■

Proposition A.4 shows that, when collusion is ex-ante, the principal finds it optimal to let the monitor and the agent collude a fraction of the time. Intuitively, when collusion is ex-ante the only way in which the principal can completely deter the agent from taking action $c = 1$ is by always paying the monitor a wage $w = \pi_A/q$: if the principal pays lower wages with positive probability, the agent will make an offer $\tau \geq 0$ that a fraction of low paid monitors will accept and will take action $c = 1$ every time she faces a monitor with a sufficiently low wage. The optimal distribution in Proposition A.4 balances the cost π_P of letting the monitor and the agent collude, and the benefit of paying lower expected wages to the monitor.

B Maxmin and Bayesian optimal policies

This appendix characterizes optimal wage distributions given a fixed budget w_0 when the principal maximizes subjective expected utility, or maxmin expected utility.

B.1 Maxmin optimal policy design

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-criminal. We consider a principal who treats environment F_η, λ as a choice variable available to an adversarial Nature. We emphasize that threshold $\overline{\pi}_A$ depends on F_η and λ by using the notation $\overline{\pi}_A(F_w, F_\eta, \lambda)$.

Taking budget w_0 as fixed, we ask what is the maxmin crime-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-criminal threshold affordable under budget w_0, *when the cost of keeping an agent of type π_A non-criminal is given by the cost function $W^*(\cdot)$*, defined in Proposition 1, i.e. let $\overline{\pi}_A^0$ be the unique solution to $W^*(\overline{\pi}_A^0) = w_0$. The following result holds.

Proposition B.1 (max-min optimal incentives)**.** *The max-min optimal level of non-criminality 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 wage distribution obtained in Section 2: $F_w^*(w) = \frac{k-\overline{\pi}_A^0}{k-qw}$. Indeed, the worst case environment for the principal is also that of Section 2, i.e. it sets $F_\eta(0) = 1$ and $\lambda = 1$.

Proof of Proposition B.1. The payoff that an agent gets from being criminal 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 $F_w^* = \frac{k-\pi_A}{k-qw}$. When the principal has a budget constraint w_0, the optimal wage schedule under the worst-case environment is $F_w^* = \frac{k-\overline{\pi}_A^0}{k-qw}$, where $\overline{\pi}_A^0$ is such that $W^*(\overline{\pi}_A^0) = w_0$. ■

B.2 Bayesian optimal incentives

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

Fix a target threshold π_A for which agents will choose to be non-criminal, as well as a wage policy F_w. An agent of type π_A chooses to remain non-criminal if and only if, for all possible bribes $\tau \in [0, \pi_A]$,

$$\begin{aligned} &\pi_A - k + \lambda(k-\tau)\text{prob}(\eta + qw < \tau) \leq 0 \qquad (18) \\ \iff &\text{prob}(\eta + qw < \tau) \leq \frac{k-\pi_A}{\lambda(k-\tau)}. \end{aligned}$$

Define

$$m_0 \equiv \min_{\tau \in [0,\pi_A]} \frac{k-\pi_A}{\lambda(k-\tau)\text{prob}(\eta < \tau)} \qquad (19)$$

and denote by τ_0 the highest solution to (19). Note that agents with type π_A such that $m_0 \geq 1$ choose to remain non-criminal for any wage distribution.[22] We focus on agents of type π_A such that $m_0 < 1$.

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

Let $\overline{\tau} \equiv \frac{\pi_A-(1-\lambda)k}{\lambda}$ and note that $\overline{\tau} > \tau_0$ for all π_A such that $m_0 < 1$.[23] Denote by Φ the operator over c.d.f.s F such that for all $w \in [0, +\infty)$,

$$\Phi(F)(w) = \begin{cases} m_0 & \text{if } w \in [0, \frac{\tau_0}{q}), \\ \min\left\{1, \frac{k-\pi_A}{f_\eta(0)\lambda(k-qw)^2} - \int_0^{qw} \frac{f'_\eta(\widehat{\eta})}{f_\eta(0)} F\left(w - \frac{\widehat{\eta}}{q}\right) \mathrm{d}\widehat{\eta}\right\} & \text{if } w \in [\frac{\tau_0}{q}, \frac{\overline{\tau}}{q}), \\ 1 & \text{if } w \geq \frac{\overline{\tau}}{q}. \end{cases} \tag{20}$$

Proposition B.2 (Bayesian-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 (18) *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 B.2 echoes Proposition 1. Incentive compatibility of non-criminal behavior at every $\tau \in [0, \pi_A]$ implies a bound on the distribution of crime costs $\eta + qw$. The intuition for point (i) comes from the fact that $\text{prob}(\eta + qw < \tau) = \text{prob}(\eta < \tau)F_w(0) + \text{prob}(\eta + qw < \tau | qw \in (0, \tau))\text{prob}(qw \in (0, \tau))$. 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 $(0, \tau_0/q)$.

We begin the proof of Proposition B.2 with a few preliminary lemmas. It is useful to note that, for any wage schedule F_w, $\text{prob}(\eta + qw < \tau) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq) dF_w(w)$. Incentive constraint (18) can then be written as: for all $\tau \in [0, \pi_A]$,

$$\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq) dF_w(w) \leq \frac{k - \pi_A}{\lambda(k - \tau)}. \tag{21}$$

Note that, for an agent with type π_A and for any wage distribution F_w, (21) is satisfied for all $\tau \geq \overline{\tau} = \frac{\pi_A-(1-\lambda)k}{\lambda}$.[24] Therefore, a principal who wants to incentivize agents with

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

[24] For all $\tau \geq \overline{\tau}$, $\frac{k-\pi_A}{\lambda(k-\tau)} \geq \frac{k-\pi_A}{\lambda(k-\overline{\tau})} = 1 \geq \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq) dF_w(w)$.

type $\pi'_A \leq \pi_A$ to be non-criminal will never find it optimal to pay wages larger than $\frac{\overline{\tau}}{q}$.[25] Therefore, 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 (21) for all $\tau \in [0,\overline{\tau}]$ and $F_w(0) < m_0$, there exists a distribution $\tilde{F}_w$ which also satisfies (21) for all $\tau \in [0,\overline{\tau}]$ such that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$.*

Proof. Let F_w be a wage schedule that satisfies (21) for all $\tau \in [0,\overline{\tau}]$ with $F_w(0) < m_0$. Suppose first that F_w is such that (21) is satisfied with slack for all $\tau \in [0,\overline{\tau}]$. Fix $\gamma > 0$ and let $\tilde{F}_w$ be a distribution such that for all $w \geq 0$, $\tilde{F}_w(w) = \min\{F_w(w) + \gamma, 1\}$. Clearly, $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. Moreover, since (21) is satisfied with slack for all τ under F_w, by choosing γ small we can guarantee that (21) is satisfied for all τ under $\tilde{F}_w$.

Suppose next that F_w is such that (21) binds for some offer τ. Let $\hat{\tau}$ be the lowest τ at which (21) binds, so that $\int_0^{\frac{\hat{\tau}}{q}} F_\eta(\hat{\tau} - wq)dF_w(w) = \frac{k-\pi_A}{\lambda(k-\hat{\tau})}$. 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 $\int_0^{\frac{\hat{\tau}}{q}} F_\eta(\hat{\tau} - wq)dF_w(w) = F_w(0)F_\eta(\hat{\tau}) = \frac{k-\pi_A}{\lambda(k-\hat{\tau})}$, 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 (19)).

We construct an alternative wage distribution $\hat{F}_w$ as follows. Fix $\gamma \in (0, F_w(\frac{\hat{\tau}}{q}) - F_w(0))$ and let $\hat{F}_w$ be such that: (i) $\hat{F}_w(w) = F_w(0) + \gamma$ for all $w \in [0, \frac{\hat{\tau}}{q}]$, (ii) $\hat{F}_w(w) = F_w(w) - (F_w(\frac{\hat{\tau}}{q}) - \hat{F}_w(\frac{\hat{\tau}}{q})) = F_w(w) - (F_w(\frac{\hat{\tau}}{q}) - F_w(0) - \gamma)$ for all $w \in (\frac{\hat{\tau}}{q}, \frac{\overline{\tau}}{q})$ and (iii) $\hat{F}_w(\frac{\overline{\tau}}{q}) = 1$. Note that $\hat{F}_w$ is a transformation of F_w that shifts γ of the mass that F_w has in $(0, \frac{\hat{\tau}}{q}]$ to 0 and the remaining $F_w(\frac{\hat{\tau}}{q}) - F_w(0) - \gamma$ to $\frac{\overline{\tau}}{q}$. By choosing γ small we can guarantee that (21) is satisfied for all $\tau \in [0,\hat{\tau}]$ under $\hat{F}_w$.

[25] To see this, let F_w be a wage profile that satisfies (21) for all τ, with $F_w(\frac{\overline{\tau}}{q}) < 1$. Let $\tilde{F}_w$ be such that $\tilde{F}_w(w) = F_w(w)$ for $w < \frac{\overline{\tau}}{q}$ and $\tilde{F}_w(w) = 1$ for $w \geq \frac{\overline{\tau}}{q}$. Clearly, $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. Moreover, for all $\tau < \overline{\tau}$, $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\tilde{F}_w(w) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)dF_w(w) \leq \frac{k-\pi_A}{\lambda(k-\tau)}$, so $\tilde{F}_w$ also satisfies (21) for all τ.

We now show that (21) is satisfied for all $\tau > \hat{\tau}$ under $\hat{F}_w$. Note first that for all $\tau \in [\hat{\tau}, \overline{\tau})$

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

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 further 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 (21) binds at $\hat{\tau}$ under F_w and the inequality follows since (21) is satisfied at $\hat{\tau}$ under $\hat{F}_w$. Since (21) is satisfied for all τ under F_w, $\frac{k-\pi_A}{\lambda(k-\tau)} \geq \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)dF_w(w) > \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w)$ for all $\tau \in (\hat{\tau}, \overline{\tau})$; that is, (21) 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 $\tilde{F}_\varepsilon(w) = \hat{F}_w(w)$ for all $w < \frac{\overline{\tau}-\varepsilon}{q}$ and $\tilde{F}_\varepsilon(w) = \hat{F}_w(w) + (F_w(\frac{\hat{\tau}}{q}) - F_w(0) - \gamma)$ for all $w \in [\frac{\overline{\tau}-\varepsilon}{q}, \frac{\overline{\tau}}{q})$; i.e., $\tilde{F}_\varepsilon$ is a transformation of $\hat{F}_w$ that puts the mass that $\hat{F}_w$ has on $\frac{\overline{\tau}}{q}$ at $\frac{\overline{\tau}-\varepsilon}{q}$. 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 $\varepsilon > 0$, (21) is satisfied for all $\tau \leq \overline{\tau} - \varepsilon$ under $\tilde{F}_\varepsilon$. On the other hand, for all $\tau \in (\overline{\tau} - \varepsilon, \overline{\tau})$, $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\tilde{F}_\varepsilon(w) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq)d\hat{F}_w(w) + F_\eta(\tau - (\overline{\tau} - \varepsilon))(F_w(\frac{\hat{\tau}}{q}) - F_w(0) - \gamma)$ is continuous and increasing in ε. Since (21) holds with slack for all $\tau \in (\hat{\tau}, \overline{\tau})$ under $\hat{F}_w$, for ε small (21) also holds for all τ under $\tilde{F}_\varepsilon$.

Let $\overline{\varepsilon} \equiv \sup\{\varepsilon : (21) \text{ holds for all } \tau \in [0, \overline{\tau}] \text{ under } \tilde{F}_\varepsilon\}$ and let $\tilde{F}_w = \tilde{F}_{\overline{\varepsilon}}$.[26] Note that there must exist $\tau' > \overline{\tau} - \overline{\varepsilon}$ such that (21) holds with equality at τ' under $\tilde{F}_w$; i.e., such that

$$\frac{k - \pi_A}{\lambda(k - \tau')} = \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq)d\tilde{F}_w(w) \geq \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq)dF_w(w), \tag{22}$$

where the inequality follows since (21) holds for all τ under F_w.

We now use (22) to show that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. Note that $\tilde{F}_w$ is a transformation of F_w

[26]For $\varepsilon = \overline{\tau}$, the cdf $\tilde{F}_\varepsilon = \tilde{F}_{\overline{\tau}}$ is such that $\tilde{F}_{\overline{\tau}}(w) = F_w(\hat{\tau}/q)$ for all $w \in [0, \hat{\tau}/q]$ and $\tilde{F}_{\overline{\tau}}(w) = F_w(w)$ for all $w > \hat{\tau}/q$. Since (21) holds with equality at $\hat{\tau}$ under F_w, (21) is not satisfied under $\tilde{F}_{\overline{\tau}}$. Hence, $\overline{\varepsilon} < \overline{\tau}$.

that shifts some of the mass that F_w has on $[0, \frac{\hat{\tau}}{q}]$ to 0 and the rest of this mass to $\frac{\overline{\tau}-\overline{\varepsilon}}{q}$. Since F_η is strictly concave, (22) implies that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$; otherwise, if $\mathbb{E}_{\tilde{F}_w}[w] \geq \mathbb{E}_{F_w}[w]$ the distribution of $\tau' - wq$ under wage schedule F_w would second-order stochastically dominate the distribution of $\tau' - wq$ under wage schedule $\tilde{F}_w$, and so (22) would not hold. ■

Lemma B.2. *Suppose F_w is such that $F_w(0) = m_0$. If (21) 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. Suppose by contradiction that $F_w(w) > F_w(0) = m_0$ for $w < \frac{\tau_0}{q}$. Then, $\int_0^{\frac{\tau_0}{q}} F_\eta(\tau_0 - wq)dF_w(w) > m_0 F_\eta(\tau_0) = \frac{k-\pi_A}{\lambda(k-\tau_0)}$, and so (21) 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 (21) for all τ. If F_w is such that (21) 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 (21) for all τ such that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$.*

Proof. Suppose that there is an interval $(\tau_1, \tau_2) \subset [\tau_0, \overline{\tau}]$ such that (21) is satisfied with slack for all $\tau \in (\tau_1, \tau_2)$ under F_w, with $F_w(\frac{\tau}{q}) < 1$ for all $\tau \in (\tau_1, \tau_2)$. There are two possibilities: (i) (21) does not bind for all $\tau > \tau_1$, or (ii) (21) binds at some $\tilde{\tau} \geq \tau_2$. 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 $\tilde{F}_w(w) = F_w(w)$ for all $w < \overline{w} - \gamma$, and $\tilde{F}_w(\overline{w} - \gamma) = 1$. Clearly, $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$. Since (21) is satisfied with slack for all $\tau > \tau_1$ under policy F_w, for γ small enough (21) is also satisfied for all τ under $\tilde{F}_w$.

Consider next case (ii). Without loss of generality, assume that (21) 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 $\tau \in (\tau_1, \hat{\tau}]$, (iii) $\hat{F}_w(\frac{\tau}{q}) = \hat{F}_w(\frac{\hat{\tau}}{q})$ for all $\tau \in (\hat{\tau}, \tau_2]$, (iv) $\hat{F}_w(\frac{\tau}{q}) = F_w(\frac{\tau}{q}) - (F_w(\frac{\tau_2}{q}) - \hat{F}_w(\frac{\tau_2}{q}))$ for all $\tau \in (\tau_2, \overline{\tau})$, and (v) $\hat{F}_w(\frac{\overline{\tau}}{q}) = 1$. Note that $\hat{F}_w$ is a transformation of F_w that shifts $\gamma(\hat{\tau} - \tau_1)$

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 the mass that F_w has over $[\frac{\tau_1}{q}, \frac{\tau_2}{q}]$ to $\frac{\overline{\tau}}{q}$. Since (21) is slack over (τ_1, τ_2) under F_w, there exists γ and $\hat{\tau} \in (\tau_1, \tau_2)$ such that (21) is satisfied for all $(\tau_1, \tau_2]$ under $\hat{F}_w$. Moreover, since $\hat{F}_w(w) = F_w(w)$ for all $w \leq \frac{\tau_1}{q}$, (21) is satisfied for all $\tau \leq \tau_1$ under $\hat{F}_w$.

We now show that (21) 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 (21) binds at τ_2 under F_w and the inequality follows since (21) is satisfied at τ_2 under $\hat{F}_w$. Since (21) 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, (21) 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 last part of the proof of Lemma B.1. For each $\varepsilon > 0$, let $\tilde{F}_\varepsilon$ be such that $\tilde{F}_\varepsilon(w) = \hat{F}_w(w)$ for all $w < \frac{\overline{\tau} - \varepsilon}{q}$ and $\tilde{F}_\varepsilon(w) = \hat{F}_w(w) + F_w(\frac{\tau_2}{q}) - \hat{F}_w(\frac{\tau_2}{q})$ for all $w \geq \frac{\overline{\tau} - \varepsilon}{q}$; i.e., $\tilde{F}_\varepsilon$ is a transformation of $\hat{F}_w$ that moves the mass that $\hat{F}_w$ puts at $\frac{\overline{\tau}}{q}$ to $\frac{\overline{\tau} - \varepsilon}{q}$. Note that $\int_0^{\frac{\tau}{q}} F_\eta(\tau - wq) d\tilde{F}_\varepsilon(w) = \int_0^{\frac{\tau}{q}} F_\eta(\tau - wq) d\hat{F}_w(w) \leq \frac{k-\pi_A}{\lambda(k-\tau)}$ for all $\tau \leq \overline{\tau} - \varepsilon$. Therefore, for all $\varepsilon > 0$, (21) holds for all $\tau \leq \overline{\tau} - \varepsilon$ under $\tilde{F}_\varepsilon$. Moreover, since (21) holds with slack for all $\tau \in (\tau_2, \overline{\tau})$ under $\hat{F}_w$, for ε small (21) also holds for all $\tau \geq \overline{\tau} - \varepsilon$ under $\tilde{F}_\varepsilon$.

Let $\overline{\varepsilon} \equiv \sup\{\varepsilon : (21) \text{ holds for all } \tau \in [0, \overline{\tau}] \text{ under } \tilde{F}_\varepsilon\}$ 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 (21) holds with equality at τ' under $\tilde{F}_w$; that is, such that

$$\frac{k - \pi_A}{\lambda(k - \tau')} = \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq) d\tilde{F}_w(w) \geq \int_0^{\frac{\tau'}{q}} F_\eta(\tau' - wq) dF_w(w), \tag{23}$$

where the inequality follows since (21) 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 $[\frac{\tau_1}{q}, \frac{\hat{\tau}}{q}]$ and the rest to $\frac{\overline{\tau}-\bar{\varepsilon}}{q}$. Since F_η is strictly concave, (23) implies that $\mathbb{E}_{\tilde{F}_w}[w] < \mathbb{E}_{F_w}[w]$; otherwise, if $\mathbb{E}_{\tilde{F}_w}[w] \geq \mathbb{E}_{F_w}[w]$ the distribution of $\tau' - wq$ under policy F_w would second-order stochastically dominate the distribution of $\tau' - wq$ under policy $\tilde{F}_w$, and (23) would not hold. ■

We can finally turn to Proposition B.2 itself.

Proof of Proposition B.2. 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 (21) 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 $F_w^*(w) < 1$,

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

It follows that the optimal distribution F_w^* is the solution to $F_w^* = \Phi(F_w^*)$, where $\Phi(\cdot)$ is defined in (20).

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 $w \in (\frac{\tau_0}{q}, \frac{\overline{\tau}}{q})$,

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

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 $d < 1$. ■

C Proofs

C.1 Proofs for Section 2

Proof of Lemma 1. Under *collusion*, the monitor's payoff from accepting an offer τ from a criminal agent is $\tau + (1 - q)w$. Her payoff from rejecting the offer of a criminal 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 criminal agent gets under collusion is $\pi_A - k + \max_\tau (k - \tau)\text{prob}(qw < \tau)$, while her payoff from being non-criminal is 0. If the agent expects to make a bribe offer $\tau > \pi_A$, her payoff from crime is $\pi_A - k + (k - \tau)\text{prob}(qw < \tau) < 0$, so she would strictly prefer to be non-criminal. ■

C.2 Proofs for Section 3

Proof of Lemma 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 criminal. Note that a criminal 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 an offer $\tau > k$.

Finally, when the agent is non-criminal, 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-criminal 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 over $[0,k]$. 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 $\tau_0 > q(w_0 + \gamma)$. For any $\epsilon \in [0,1]$, let $F_w^\epsilon = (1-\epsilon)\mathbf{1}_{w=w_0} + \epsilon F_w$; i.e., F_w^ϵ is the mixture between a deterministic wage w_0 and policy F_w. Since F_η is strictly concave over $[0,k]$, $(k-\tau)\mathbb{E}_{F_w^\epsilon}[F_\eta(\tau - qw)] < (k-\tau)F_\eta(\tau - qw_0)$ 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

$$(k-\tau_\epsilon)\mathbb{E}_{F_w^\epsilon}[F_\eta(\tau_\epsilon - qw)] < (k-\tau_\epsilon)F_\eta(\tau_\epsilon - qw_0) \leq (k-\tau_0)F_\eta(\tau_0 - qw_0),$$

where the last inequality follows since τ_0 solves $\max_\tau (k-\tau)F_\eta(\tau - qw_0)$. It follows that for ϵ small the expected payoff a criminal agent obtains under F_w^ϵ is strictly smaller than the one she obtains 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 criminal 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 criminal 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,

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

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 crime compared to deterministic wage w_0. ■

C.3 Proofs for Section 4

Proof of Lemma 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,

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

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 $\overline{R}_0 = 0$ is

$$k - qw_0 - \overline{\eta} > (k - qw_0)(1-p) \iff k - qw_0 > \frac{\overline{\eta}}{p}. \tag{24}$$

This condition expresses that it is optimal for the agent to offer a bribe $\tau = qw_0 + \overline{\eta}$ rather than $\tau = qw_0$ under the deterministic wage w_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

$$k - q\overline{w} - \overline{\eta} < (k - q\overline{w})(1-p) \iff k - q\overline{w} < \frac{\overline{\eta}}{p}. \tag{25}$$

Since $\overline{w} > w_0$, it is immediate that for any ϵ, one can find values $p, \overline{\eta} < \epsilon$, such that conditions (24) and (25) 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 (24) 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. ■

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.

Basu, K. (2011): "Why, for a Class of Bribes, the Act of Giving a Bribe should be Treated as Legal," .

Basu, K., K. Basu, and T. Cordella (2014): "Asymmetric punishment as an instrument of corruption control," *World Bank Policy Research Working Paper*.

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

Bergemann, D., B. Brooks, and S. Morris (2015): "The Limits of Price Discrimination," *The American Economic Review*, 105.

Bergemann, D. and M. Pesendorfer (2007): "Information structures in optimal auctions," *Journal of Economic Theory*, 137, 580–609.

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," .

Burguet, R. and Y.-K. Che (2004): "Competitive procurement with corruption," *RAND Journal of Economics*, 50–68.

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.

Carroll, G. (2013): "Robustness and Linear Contracts," *Stanford University Working Paper*.

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., D. Condorelli, and J. Kim (2013): "Weak Cartels and Collusion-Proof Auctions," .

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

——— (2009): "Optimal collusion-proof auctions," *Journal of Economic Theory*, 144, 565–603.

Condorelli, D. and B. Szentes (2016): "Buyer-Optimal Demand and Monopoly Pricing," Tech. rep., Mimeo.

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*.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

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

Khan, A. Q., A. I. Khwaja, and B. A. Olken (2014): "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors," .

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.

Pavlov, G. (2008): "Auction design in the presence of collusion," *Theoretical Economics*, 3, 383–429.

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

Punch, M. (2009): *Police corruption: Deviance, accountability and reform in policing*, Routledge.

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.

Segal, I. (2003): "Optimal pricing mechanisms with unknown demand," *The American economic review*, 93, 509–529.

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.