Corruption, Intimidation and Whistleblowing: A Theory of Optimal Intervention

Sylvain Chassang
Princeton University

Gerard Padró i Miquel*
London School of Economics

April 23, 2012.

Preliminary and Incomplete

Abstract

We consider a game between a principal, an agent, and a monitor in which the principal would like to rely on messages by the monitor to target intervention against a misbehaving agent. The difficulty is that the agent can credibly threaten to retaliate against likely whistleblowers in the event of intervention. Taking information as given, intervention policies that are more responsive to the monitor's messages provide greater incentives for the agent to behave well. However, making intervention responsive to the monitor's message also facilitates retaliation by corrupt agents and limits information provision. Indeed very responsive policies lead to silent corruption in which the agent dissuades any reporting. Therefore, successful intervention policies necessarily garble the information provided by the monitor. Furthermore, we show that effective intervention policies can be identified using only messaging data, and that to robustly rule out silent corruption it is necessary to experience situations where corruption is taking place and is being reported.

Keywords: corruption, monitoring, whistleblowing, threats, retaliation, optimal intervention.

*Chassang: chassang@princeton.edu, Padró i Miquel: g.padro@lse.ac.uk.

1 Introduction

This paper explores anti-corruption mechanisms in which a principal relies on messages by an informed monitor to target intervention against a potentially misbehaving agent. The difficulty is that the agent can credibly threaten to retaliate against likely whistleblowers. We show that taking information as given, intervention policies that are more responsive to the monitor's messages provide greater incentives for the agent to behave well. However, making intervention responsive to the monitor's message also facilitates retaliation by corrupt agents and limits endogenous information provision. As a consequence there is a trade-off between eliciting information and using that information. This makes finding effective intervention policies difficult: imagine that no complaints are received, does this mean that there is no underlying corruption, or does it mean that would-be whistleblowers are being silenced by threats and intimidation? We investigate optimal intervention patterns and suggest ways to identify effective intervention strategies using limited data.

Our framework encompasses various forms of corruption: bribe collection by state officials, collusion between police officers and organized crime, fraud by sub-contractors in public good projects, breach of fiduciary duty by a firm's top executives and so on... Retaliation can also take several forms: an honest bureaucrat may be socially excluded by his colleagues and denied promotion; police officers suspected of collaborating with Internal Affairs may have their life threatened by lack of prompt support[1], whistleblowers may be harrassed or see their careers derailed. In many cases retaliation is facilitated by the fact that only a few colleagues, subordinates, or frequent associates are informed about the agent's misbehavior. However, group punishments may also be used, for instance, entire communities may be denied access to public services.[2] The key ingredients of our analysis are that: (1) there is significant information about corrupt agents which the principal wants to obtain;

[1] See Punch (2009) for examples of punishment of informants in a study of police corruption.

[2] See Ensminger (2012) for an example of a whole village being cut out from the water supply by corrupt bureaucrats retaliating against complaints.

(2) the individuals who have this information or are able to pass it on to the principal can be punished by the agent.

The model considers a dynamic game played by a principal, an agent and a single monitor. Both the principal and the agent have commitment power, and they act sequentially. The principal first commits to an intervention strategy as a function of the information obtained from the monitor, i.e. to a likelihood of intervention as a function of messages "corrupt" and "non-corrupt". The agent then commits to a retaliation strategy against the monitor as a function of what happens to him. Finally, the monitor observes the corruption behavior of the agent and chooses what message to send to the principal. A key element of our modelling approach is to recognize that the principal need not have full control over the agent's and the monitor's outcomes following intervention. For instance, a principal may decide to sue the agent, but the agent's final outcome could be determined by an exogenous judiciary process. Similarly, whistleblower protection schemes may not fully shield the monitor against indirect punishments such as ostracism, or harassment. Furthermore we do not assume that the monitor necessarily desires intervention against corrupt agents. For instance, in a development context, corruption scandals frequently lead to a withdrawal of funding by NGOs or the World Bank.[3] Since 10% on the dollar is better than 0% on the dollar, citizens may prefer not to complain about a corrupt sub-contractor. It can also be the case that the monitor benefits from having an honest agent investigated. For instance, corruption investigations may be used to weaken competing colleagues.

Our analysis emphasizes two sets of results. We first show that optimal intervention strategies must necessarily garble the information provided by the agent. Otherwise the principal's behavior is a perfect signal of the monitor's message, which makes retaliation very effective. In particular, the likelihood ratio of intervention rates under messages "corrupt" and "not corrupt" must be bounded above. An immediate consequence is that in equilibrium, the likelihood of intervention against non-corrupt agents must be bounded away from zero.

[3]See Ensminger (2012).

Furthermore, it may be optimal not to intervene against agents known to be corrupt. The reason for this is that reducing intervention on the corrupt allows to reduce costly intervention on non-corrupt agents without increasing the ratio of intervention rates.

The second set of results characterizes the geometry of messaging and corruption decisions as a function of intervention rates. We show that the region of the intervention-strategy space in which corruption occurs is star-shaped around the origin. Furthermore, keeping corruption behavior constant, messages by agents depend only on the ratio of intervention rates. We show that this allows to identify whether silent corruption is occurring—i.e. corruption which the monitor chooses not to report—using only complaint data.

This paper hopes to contribute to a growing effort to understand the effectiveness of counter-corruption measures. In recent years, the World Bank, the OECD and the United Nations have launched new initiatives to improve governance, in the belief that a reduction in corruption can improve the growth trajectory of developing countries.[4] Growing micro-economic evidence confirms the importance of corruption on public service provision and public expenditure in education or health (see Olken and Pande (2011) for a recent review), and suggests that appropriate incentive design can effectively reduce misbehavior (Olken (2007); Duflo et al. (forthcoming)). While relatively unstudied, retaliation has been recognized as playing a significant role in sustaining corrupt agents. For instance, in the words of Banerjee and Duflo (2006) "the beneficiaries of education and health services are likely to be socially inferior to the teacher or health care worker, and a government worker may have some power to retaliate against them."

This work also contributes to the literature on monitoring in principal-agent relationships. Rahman (forthcoming) formalizes the idea of "mystery shoppers" and emphasizes the value of random message-based incentives to jointly incentivize effort provision by the agent and

[4]See Mauro (1995) for an early paper highlighting the association of corruption and lack of growth. Shleifer and Vishny (1993) and Acemoglu and Verdier (1998, 2000) all provide theories of corruption that introduce distortions above and beyond the implicit tax that corruption imposes.

the monitor supposed to evaluate the agent. The role of mixed strategies in our work is entirely different: monitoring is costless and randomization occurs only to garble the information content of the principal's intervention behavior.[5] Unlike Tirole (1986) or Laffont and Martimort (1997), we do not investigate the possibility of collusion between the agent and the monitor. Instead we focus on retaliation rather than side payments as a mean for the agent to discipline a potential monitor. Indeed, threats figure prominently as a mean to dissuade communication, both as a complement and a substitute to side-payments. A reason for this is that punishment costs often need not be paid on the equilibrium path. Furthermore, our basic insight that the principal's strategy should garble the information provided by the monitor to make incentive provision by the agent more difficult remains true regardless of the incentive tools that the agent can use. Finally our work shares much of its motivation with recent work on privacy in mechanism design (Izmalkov et al., 2011; Ghosh and Roth, 2010; Nissim et al., 2011; Gradwohl, 2012).

The paper is structured as follows: Section 2 describes the model; Section 3 delineates the interplay between intervention, retaliation and information provision in a simple context; Section 4 provides a general characterization of corruption and messaging patterns as a function of the intervention profile; Section 5 describes optimal intervention patterns as a function of the monitor's preference for or against intervention; Section 6 discusses identification issues; Section 7 concludes.

2 Framework

Players, timing and actions. There are three players, which we refer to as the principal P, the agent A and the monitor M.[6] The timing of actions in the game are as follows.

1. The agent chooses whether to be corrupt ($c = 1$) or not ($c = 0$). The monitor observes

[5] Eeckhout et al. (2010) propose a different theory of optimal random intervention based on non-linear responses of criminal behavior to the likelihood of enforcement.

[6] Throughout the paper we refer to the principal and the monitor as she, and to the agent as he.

corruption c and sends a message $m \in \{0, 1\}$ to the principal.

2. The principal observes the monitor's message m and triggers an intervention or not: $i \in \{0, 1\}$. Intervention has uncertain payoff-relevant consequences $z = (z_A, z_P, z_M) \in Z \subset \mathbb{R}^k$ for the agent, the principal and the monitor.

3. The agent retaliates with intensity $r \in [0, +\infty)$ against the monitor.

Observables, consequences, and payoffs. The monitor costlessly observes the agent's corruption decision $c \in \{0, 1\}$. In contrast, the principal does not observe this decision. The act of triggering an intervention, $i \in \{0, 1\}$, is observed by all players. Intervention by the principal generates consequences $z = (z_A, z_P, z_M) \in Z$ for all players, where these consequences can be uncertain. Player $x \in \{A, P, M\}$ observes only her own consequences, z_x. The distribution of consequences z is denoted by $f(z|c, m)$.

Note that the distribution of consequences $f(z|c, m)$ depends both on corruption status and message m. In other words, the distribution of outcomes f subsumes all exogenous or endogenous dependency of outcomes on messages. This reduced-form abstraction allows us to focus on a simple intervention decision common to a variety of settings where the principal may or may not have full control over the consequences of intervention. This includes settings in which actual punishment of corrupt agents is performed by an external entity such as a law-enforcement agency, or settings where intervention attracts media interest and affects the reputation of those involved. This formalism also allows for consequences to contain information regarding the message sent by the monitor. For instance a corrupt agent may be able to obtained leaked information about likely whistleblowers. Note that this exogenous dependency of outcomes on messages means that the revelation principle need not hold since messages here have institutional meaning beyond the control of the principal. Finally, this setting also allows for the investigations or audits triggered by intervention to fail to determine the existence of corruption with full certainty.

As a function of $c \in \{0,1\}, i \in \{0,1\}, z \in Z$ and $r > 0$, realized payoffs u_A, u_P and u_M are described by

$$u_A = \pi_A \times c + i \times v_A(z_A) - k_A(r)$$
$$u_P = \pi_P \times c + i \times v_P(z_P) - k_P(r)$$
$$u_M = \pi_M \times c + i \times v_M(z_M) - r.$$

We normalize $z = (0,0,0)$ and $v_A(0) = v_P(0) = v_M(0) = 0$ if no intervention is triggered, and $k_A(0) = k_P(0) = 0$ if no retaliation occurs. We make the following assumptions regarding the structure of payoffs.

Assumption 1. *Corruption payoffs π_A and π_P satisfy $\pi_A > 0$ and $\pi_P < 0$.*
Retaliation costs $k_A(r)$ and $k_P(r)$ are weakly positive, and weakly increasing in r.
Expected payoffs conditional on intervention ($i = 1$) satisfy

$$\begin{aligned}
\textit{agent payoff simplification:} \quad & \forall c \in \{0,1\}, \mathbb{E}(v_A|c, m=1) = \mathbb{E}(v_A|c, m=0) \leq 0 \\
\textit{dissuasive intervention:} \quad & \pi_A + \mathbb{E}(v_A|c=1) < \mathbb{E}(v_A|c=0) \\
\textit{costly intervention:} \quad & \forall m \in \{0,1\}, \mathbb{E}(v_P|c=0, m) \leq 0 \\
\textit{marginal preference for the truth:} \quad & \forall c \in \{0,1\}, \mathbb{E}(v_M|c, m \neq c) < \mathbb{E}(v_M|c, m=c)
\end{aligned}$$

These assumptions put some constraints on the environments we examine. First we simplify the agent's problem by assuming that the agent's payoff depends only on his intervention status. Second, the threat of certain intervention is sufficient to dissuade the agent from being corrupt. Third, intervention is costly for the principal, at least if the agent is non-corrupt: the principal wants to minimize intervention. Fourth, taking intervention as given the monitor is better off if she conveyed the truth to the principal. This restrictive assumption is meant to capture potential rewards and punishments doled out by the principal or the agency called in during the intervention, as well as non-monetary or moral

rewards that the monitor might feel by participating in the punishment of a corrupt agent. Importantly we do not make assumptions regarding the monitor's preferences for or against intervention, i.e. $\mathbb{E}(v_M|c=1,m)$ may be positive or negative.

Strategies and commitment. Both the principal and the agent can commit to strategies ex ante. The principal acts as a first mover and commits to an intervention policy $\sigma : m \in \{0,1\} \mapsto \sigma_m \in [0,1]$, where $\sigma_m = \text{prob}(i=1|m)$ is the likelihood of intervention given message m. Knowing the principal's strategy σ, the agent takes a corruption decision $c \in \{0,1\}$ and commits to a retaliation policy $r : z_A \in Z_A \mapsto r(Z_A) \in [0,+\infty)$ as a function of the consequences z_A that he observes. The monitor moves last and takes an optimal messaging decision given the commitments of both the principal and the agent. We are interested in characterizing the policy σ that guarantees the principal the highest expected utility.

3 The Trade-off Between Eliciting and Using Information

Trade-offs between eliciting and using information occur systematically in mechanism design. Here information usage is restricted by the fact that it facilitates incentive provision by a corrupt agent. This section delineates the mechanics of intervention, corruption and communication in a simple setting.

In this section, we consider the case where the agent gets no information about the monitor's message except that which is conveyed through the principal's intervention. We temporarily impose that: (1) the agent does not retaliate in the absence of intervention, and he chooses a single retaliation level r conditional on intervention $i=1$; (2) intervention profiles are such that $\sigma_1 \geq \sigma_0$ (i.e. intervention is more likely following message "corrupt" than

"non-corrupt"); (3) the monitor always sends truthful messages in the absence of retaliation; (4) intervention is costless for non-corrupt agents ($\mathbb{E}[v_A|c=0]=0$), and the principal does not care about retaliation ($k_P(r)=0$).[7] We solve the game by backward induction.

Truthtelling. Our preliminary assumptions guarantee that if the agent chooses $c=0$, there will be no retaliation and the agent will send truthful message $m=0$. Imagine now that under an intervention profile $\sigma=(\sigma_0,\sigma_1)$, the agent makes corruption decision $c=1$. The monitor will be truthful (i.e. send message $m=1$) if and only if

$$\sigma_1(\mathbb{E}[v_M|c=1,m=1]-r)\geq\sigma_0(\mathbb{E}[v_M|c=1,m=0]-r).$$

The highest value of r for which this holds is

$$\hat{r}=\frac{\sigma_1\mathbb{E}[v_M|c=1,m=1]-\sigma_0\mathbb{E}[v_M|c=1,m=0]}{\sigma_1-\sigma_0}. \tag{1}$$

Note that $\hat{r}$ is decreasing in the ratio σ_1/σ_0: when the information content of intervention is large, lower retaliation levels are required to shut-down truthful communication .

Information manipulation. We now examine the agent's incentives to shut down information channels or not, conditional on being corrupt. Since retaliation r is costly to the agent, he either picks $r=0$ and lets the monitor send truthful messages, or picks $r=\hat{r}$ and shuts-down information at the lowest possible cost. Hence, the agent will manipulate messages through the threat of retaliation if and only if:

$$\sigma_1\mathbb{E}[v_A|c=1]\leq\sigma_0(\mathbb{E}[v_A|c=1]-k_A(\hat{r})). \tag{2}$$

[7]These assumptions will either be relaxed or become results in Section 4.

Corruption. It follows from the analysis so far that the agent will choose not to be corrupt if and only if

$$\pi_A + \max\{\sigma_1 \mathbb{E}[v_A|c=1], \sigma_0(\mathbb{E}[v_A|c=1] - k_A(\hat{r}))\} \leq 0. \tag{3}$$

Note that when the agent chooses to manipulate information, increasing baseline intervention rate σ_0 reduces the gains from corruption through multiple channels:

1. it increases the expected cost of intervention $\sigma_0 \mathbb{E}[v_A|c=1]$ in equilibrium;

2. it increases the likelihood that costly retaliation happens on the equilibrium path;

3. it increases the level $\hat{r}$ of retaliation needed to shut-down communication.

In turn, increasing σ_0 has no effect on corruption when the agent chooses not to manipulate information. Therefore the overall effect of increasing baseline intervention rates is unambiguous: it reduces corruption (at least weakly).

Inversely the effect of increasing intervention rate σ_1 following message $m = 1$ is ambiguous:

1. on the one hand it increases the expected cost of intervention $\sigma_1 \mathbb{E}[v_A|c=1]$ when the agent does not manipulate information;

2. on the other hand it *decreases* the level of retaliation $\hat{r}$ needed to shut-down communication when the agent chooses to manipulate communication.

Optimal intervention. The principal's payoff is maximized either by setting $\sigma_0 = \sigma_1 = 0$ and tolerating corruption, or by finding the minimum value σ_0 such that there exist σ_1 for which (3) holds. This is achieved by the profile (σ_0, σ_1) such that (2) and (3) hold with equality (see Figure 1). Equality in (2) and (3) defines straight lines whose intersection define the optimal intervention profile. Inspection shows that it must satisfy the following properties:

1. $\sigma_0 > 0$, i.e. there will be intervention against some agents known in equilibrium not to be corrupt; otherwise the agent could commit to large but costless threats that would silence the monitor;

2. $\sigma_1 < 1$, i.e. there will not be intervention against all agents known in equilibrium to be corrupt; this allows to limit baseline intervention rate σ_0 without increasing the likelihood ratio σ_1/σ_0.

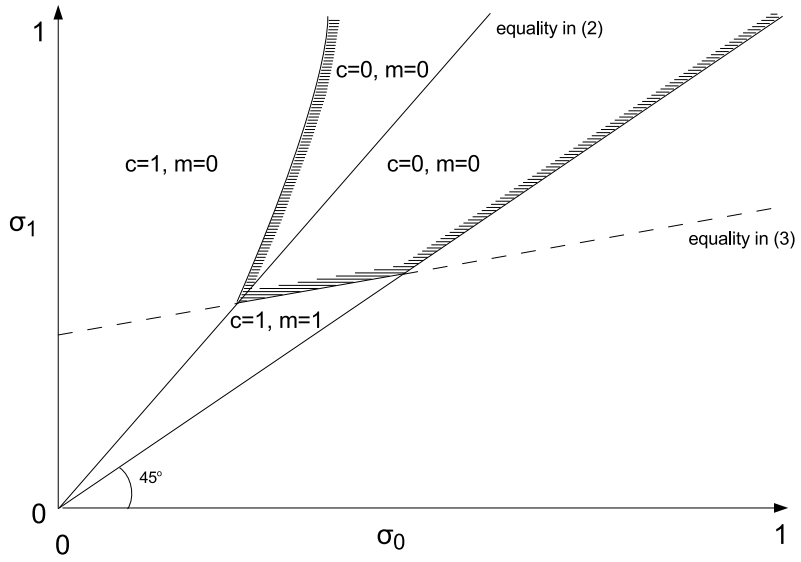


Figure 1: information and corruption as a function of intervention profiles

4 General Analysis

We now relax the restrictive assumptions made in Section 3. Section 4.1 provides basic results that simplify further analysis and establish the necessity of garbling the information content of the monitor's messages. Section 4.2 clarifies the structure of information manipulation and corruption decisions.

4.1 Preliminary results

A benchmark case. To frame the analysis it is useful to consider first the case where the monitor exogenously sends truthful messages. More specifically, assume that the monitor is an automaton with strategy $m(c) = c$. In this setting, the following lemma is immediate.

Lemma 1. *Under the assumption of exogenously truthful messages, the optimal intervention policy is to set* $\sigma_0 = 0$, *and* $\sigma_1 = 1$. *On the equilibrium path, there is no corruption, no intervention and no retaliation.*

This result follows from Assumption 1, which ensures that the agent refrains from corruption if intervention occurs with probability 1. The principal intervenes if and only if the agent is corrupt: intervention is fully responsive to the monitor's message. This will not be the case when information is endogenous.

Endogenous information. We now consider the case where the monitor's decision to send messages must be incentive compatible in the face of potential retaliation from the agent. The following lemma shows that intervention can no longer fully respond to messages.

Lemma 2 (necessary garbling)**.** *If* $|\log(\sigma_1/\sigma_0)| = +\infty$ *the agent will choose to be corrupt and threaten to retaliate upon intervention. On the equilibrium path message* $m = 0$ *will be sent, and no intervention or retaliation will occur.*

If $\sigma_0 = 0$ and the agent can induce the monitor to always send message $m = 0$, interventions do not occur along the equilibrium path. This means that the agent can commit to arbitrarily high levels of retaliation in response to interventions, and this will ensure that the monitor always sends message $m = 0$. In this way the agent can obtain the monitor's silence at no cost.

To increase the cost of silence for the agent, the principal must therefore ensure that retaliation takes place along the equilibrium path. To do so, it must commit to set up a

baseline probability of intervention σ_0 strictly greater than 0. This forces the agent to incur the costs of the promised retaliation with positive probability. The basic tension is that σ_0 increases at the same time the agent's costs of inducing silence and the principal's expected costs due to intervention. For this reason the principal might give up the fight against corruption. If the principal cannot ensure absence of corruption in equilibrium the optimal policy will have no intervention on the equilibrium path.

Lemma 3. *If there is corruption in equilibrium, an optimal policy for the principal sets* $\sigma_0 = \sigma_1 = 0$. *There is no retaliation and no intervention in equilibrium.*

4.2 Patterns of information manipulation and corruption

In this subsection we characterize the behavior of agent and monitor as a function of the intervention strategy of the principal. We say that there is information manipulation when the agent induces the monitor not to report truthfully to the principal.

Information manipulation, given corruption. We first take as given the agent's corruption decision c and focus on the agent's decision to manipulate messages sent by the monitor. We begin by describing the agent's optimal retaliation policy.

Lemma 4. *(i) Given a corruption decision c, if it is optimal for the agent to commit to a non-zero retaliation profile which induces message m, then it must be that* $\sigma_m < \sigma_{\neg m}$.

(ii) Conditional on c, it is never optimal for the agent to exert any retaliation if there is no intervention.

Point (i) is intuitive. Retaliation is costly, therefore any investment in retaliation must reduce the likelihood of intervention. Point (ii) is a natural consequence: since retaliation

is meant to discourage messages that imply a higher likelihood of intervention, it is optimal for the agent to concentrate retaliation in the event of intervention.

We now introduce additional notation. For any corruption decision c, we denote by $K_{c,m}$ the expected cost of inducing message m given corruption status c, conditional on an intervention strategy. Define $\lambda_m \equiv \sigma_m/\sigma_{\neg m}$. We have that

$$\begin{aligned} K_{c,m} = & \inf_{r:Z_A\to[0,+\infty)} \int_{Z_A} k_P(r(z_A)) f(z_A|c,m)\mathrm{d}z_A \\ & \text{s.t.} \quad \sigma_m \left[\mathbb{E}(v_M|m,c) - \mathbb{E}(r|m,c)\right] \geq \sigma_{\neg m} \left[\mathbb{E}(v_M|\neg m,c) - \mathbb{E}(r|\neg m,c)\right] \\ = & \inf_{r:Z_A\to[0,+\infty)} \int_{Z_A} k_P(r(z_A)) f(z_A|c,m)\mathrm{d}z_A \\ & \text{s.t.} \quad \lambda_m \left[\mathbb{E}(v_M|m,c) - \mathbb{E}(r|m,c)\right] - \left[\mathbb{E}(v_M|\neg m,c) - \mathbb{E}(r|\neg m,c)\right] \geq 0. \end{aligned}$$

Note that the intervention strategy of the principal, $\sigma = (\sigma_0, \sigma_1)$, only enters this expression through $(\lambda_m)_{m\in\{0,1\}}$. This, and the fact that $\lambda_{\neg m} = \lambda_m^{-1}$ have a useful implication. The cost of information manipulation is only a function of the likelihood ratio of intervention, $\lambda \equiv \sigma_1/\sigma_0$, and not of the separate components of the principal's strategy. We can thus denote the cost of information manipulation by $K_{c,m}(\lambda)$.

To describe the situations in which information is manipulated, it is helpful to denote by $m^*(c,\sigma)$ the message induced by the agent given corruption decision c and intervention profile σ. Also, denote by $c^*(\sigma)$ the agent's optimal corruption decision given intervention profile σ.

Lemma 5 (manipulation)**.** *Given an intervention profile σ, and a corruption decision c, there exists a message m such that $K_{c,m}(\lambda) = 0$ and $K_{c,\neg m}(\lambda) > 0$. The agent induces a message m^* that solves*

$$\max_{m\in\{0,1\}} \sigma_m \left[\mathbb{E}(v_A|c) - K_{c,m}(\lambda)\right].$$

This lemma simply describes the decision process of the agent. In the absence of retalia-

tion, there is a message that the monitor prefers to send. Hence, inducing this message has an associated cost of 0. It is costly to induce the opposite message. To choose which message to induce, the agent simply picks the one that maximizes returns net of manipulation costs. This directly implies that, given a decision c, the message m^* induced by the agent is a function only of the likelihood ratio of intervention, $\lambda = \sigma_1/\sigma_0$.

The following theorem characterizes the behavior of agent and monitor conditional on the intervention profile σ they face.

Theorem 1 (patterns of manipulation and corruption)**.**

(i) Given a corruption decision c, for any $m \in \{0,1\}$ the set of intervention profiles in which the monitor sends the same message, $\{\sigma \mid m^*(c,\sigma) = m\}$ is a cone with origin at $(0,0)$. In other terms, pick two intervention profiles σ and $\rho\sigma$ with $\rho > 0$, it must be that $m^*(c,\sigma) = m^*(c,\rho\sigma)$.

(ii) The set of intervention profiles such that the agent's optimal behavior is to be corrupt, $\{\sigma | c^*(\sigma) = 1\}$, is star-shaped around the origin $(0,0)$. In other terms,

$$\forall \rho \in [0,1], \quad c^*(\sigma) = 1 \Rightarrow c^*(\rho\sigma) = 1,$$
$$\forall \rho \in [1,+\infty), \quad c^*(\sigma) = 0 \Rightarrow c^*(\rho\sigma) = 0.$$

Point (i) essentially corresponds to the fact that taking corruption decisions as given, manipulation decisions by the agent depend only on the likelihood ratio λ. For instance, for σ_1/σ_0 high, manipulation will be cheap on the equilibrium path. If instead σ_1/σ_0 is close to 1, manipulation will become arbitrarily costly on the equilibrium path.

The decision to be corrupt will depend on both the ratio and the level of intervention rates. Point (ii) shows that proportional increases in intervention rates always diminish corruption, while proportional decreases in intervention rates always increase corruption.

Interestingly, these patterns of manipulation can be used to identify whether unreported corruption is going on.

Corollary 1 (a message-based test against unreported corruption)**.** *Consider an intervention profile σ. If there exists $\rho \in (0,1)$ such that $m^*(\rho\sigma) \neq m^*(\sigma)$ then $c^*(\sigma) = 0$.*

This corollary exploits the fact that if the agent is willing to complain about corruption at a given likelihood ratio σ_1/σ_0, then the message that the agent is not corrupt is trustworthy at intervention profiles that share the same likelihood ratio. Note that these results hold without assumptions of payoffs beyond Assumption 1.

5 Optimal Intervention

To further characterize optimal intervention policies we need to separate different cases as a function of the preferences of the monitor. To simplify the analysis, we maintain throughout the assumption that

$$\forall c \in \{0,1\}, \quad \text{sgn}(\mathbb{E}[v_M|c,m=0]) = \text{sgn}(\mathbb{E}[v_M|c,m=1]),$$

i.e. the sign of $\mathbb{E}[v_M|c,m]$ depends only on c.[8]

5.1 Case 1: the fearful monitor

The first case of interest is the one in which

$$\forall c \in \{0,1\}, \quad \mathbb{E}(v_M|c,m) \leq 0.$$

In words the monitor dislikes intervention. In practice, this could be the case if the monitor has also participated in illicit behavior. Intervention is costly since it might uncover the whole

[8]Note that this is consistent with the monitor having marginal preferences for the truth.

corruption network. This being said, taking intervention as given she might be relatively better off by informing the principal. Another example would be that of fraud in public good projects. If corruption scandals lead to full withdrawals of funding, local communities will prefer to avoid triggering intervention, even though they may only get a small share of the resources allotted to them.

We begin with some preliminary results that simplify the analysis.

Lemma 6. *Any optimal intervention profile that induces $c = 0$, exhibits $\sigma_1 \geq \sigma_0$. At the optimum, there is no retaliation on the equilibrium path and $m^* = c^*$.*

This lemma states that if corruption is to be prevented, a message that the agent is corrupt must be associated to a higher probability of intervention. At the optimum, either the principal is able to prevent corruption from happening, or she is not. If she is, the agent is non-corrupt and has no need to retaliate since the monitor has incentives to send truthful messages. If the principal is unable to prevent corruption, Lemma 3 shows that there is neither intervention or retaliation on the equilibrium path.

It follows from this lemma that conditional on non-corruption, the principal's objective can simply be written as

$$\max_{\sigma} \sigma_0 \mathbb{E}(v_P | c = 0, m = 0)$$
$$s.t. \;\; c^*(\sigma) = 0.$$

In other words, the principal simply wants to find the lowest baseline probability of intervention σ_0 consistent with the agent being non-corrupt. As in Section 3, the principal is limited by the fact that greater responsiveness to messages reduces the agent's cost for manipulating information.

Lemma 7. *The cost $K_{1,0}(\lambda)$ of inducing message $m = 0$ for a corrupt agent is decreasing in λ. Furthermore, $K_{1,0}(1) > 0$.*

To see this more formally, note that the incentive compatibility constraint ensuring that the monitor sends message $m = 0$ can be written as

$$\lambda \left[\mathbb{E}(v_M|c = 1, m = 1) - \mathbb{E}(r|c = 1, m = 1)\right] - \left[\mathbb{E}(v_M|c = 1, m = 0) - \mathbb{E}(r|c = 1, m = 0)\right] \leq 0.$$

This becomes easier to satisfy when λ increases, since $\mathbb{E}(v_M|c = 1, m = 1) - \mathbb{E}(r|c = 1, m = 1) \leq 0$ when the constraint binds.[9]

Armed with these results, we can now move on to characterize optimal patterns of intervention.

Truthtelling. Lemma 6 allows us to focus on profiles σ above the 45^o line. A non-corrupt agent will abstain from retaliation since it induces message $m = 0$. Conditional on corruption, the agent will choose to manipulate messages if and only if

$$\begin{aligned} &\sigma_0[\mathbb{E}(v_A|c = 1) - K_{1,0}(\lambda)] \geq \sigma_1 \mathbb{E}(v_A|c = 1) \\ \iff &\mathbb{E}(v_A|c = 1)(\lambda - 1) + K_{1,0}(\lambda) \leq 0. \end{aligned} \tag{4}$$

Since $\mathbb{E}(v_A|c = 1) < 0$, $K_{1,0}(\lambda)$ is decreasing in λ, and $K_{1,0}(1) > 0$, it follows that there exists $\lambda_T \in [1, \infty)$ such that for all $\lambda \geq \lambda_T$ a corrupt agent induces message $m = 0$, and for $\lambda \leq \lambda_T$, a corrupt agent induces message $m = 1$.

Revealed Corruption. When (4) does not hold—i.e. when a corrupt agent would choose not to manipulate information—the agent will find it optimal to be corrupt if and only if

$$\pi_A + \sigma_1 \mathbb{E}(v_A|c = 1) \geq \sigma_0 \mathbb{E}(v_A|c = 0) \iff \sigma_1 \leq \frac{\mathbb{E}(v_A|c = 0)}{\mathbb{E}(v_A|c = 1)}\sigma_0 + \frac{\pi_A}{-\mathbb{E}(v_A|c = 1)}. \tag{5}$$

[9]Note that $\mathbb{E}(v_M|c = 1, m = 0)$ is negative by assumption.

Intuitively, the agent will be corrupt when σ_1 is not too high. Note that the set of profiles such that (5) holds with equality is a straight line with intercept and slope between 0 and 1:

$$\frac{\mathbb{E}(v_A|c=0)}{\mathbb{E}(v_A|c=1)} \in (0,1) \quad \text{and} \quad \frac{\pi_A}{-\mathbb{E}(v_A|c=1)} \in (0,1).$$

Unreported Corruption. In contrast, if (4) holds so that a corrupt agent would manipulate messages, the agent will be corrupt if and only if

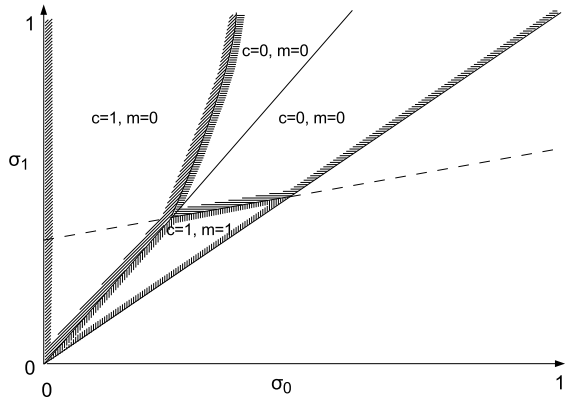
$$\begin{aligned} &\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda)] \geq \sigma_0 \mathbb{E}(v_A|c=0) \\ \iff &\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - \mathbb{E}(v_A|c=0) - K_{1,0}(\lambda)] \geq 0. \end{aligned} \tag{6}$$

Since $K_{1,0}(\lambda)$ is decreasing in λ, the boundary of corruption and non-corruption regions can be described by an increasing curve $\sigma_1(\sigma_0)$. The fact that the set of intervention profiles inducing corruption is star-shaped implies that the curve $\sigma_1(\sigma_0)$ at which the agent is indifferent between corruption and non-corruption exhibits increasing averages. Intuitively, as the baseline probability of intervention σ_0 becomes large enough, audits happen sufficiently often to dissuade corruption, even if the agent prevents the monitor from sending informative messages.

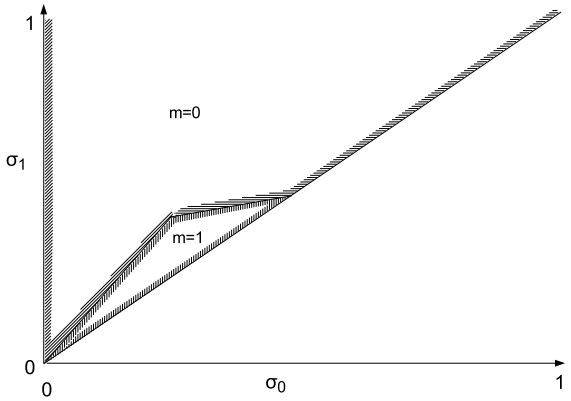
Optimum. It follows from inspection (see Figure 2 for intuition) that conditional on suppressing corruption the optimal intervention profile is such that (4) and (5) hold with equality.

Lemma 8. *The intervention profile that induces no-corruption at the minimum intervention cost is characterized by*

$$\begin{aligned} \mathbb{E}(v_A|c=1)(\lambda - 1) + K_{1,0}(\lambda) &= 0 \\ \pi_A + \sigma_1 \mathbb{E}(v_A|c=1) - \sigma_0 \mathbb{E}(v_A|c=0) &= 0. \end{aligned}$$



(a) Behavior and messages.



(b) Messages only.

Figure 2: fearful monitor

In the example illustrated by Figure (2), the optimal intervention profile is interior: σ_0 is strictly larger that 0 and σ_1 is strictly below 1. This is in fact always the case.

Corollary 2 (optimal intervention profiles are interior)**.** *The intervention profile that induces no-corruption at the minimum intervention cost is interior:*

$$0 < \sigma_0 \leq \sigma_1 < 1.$$

The fact that σ_0 must be strictly above 0 is an implication of Lemma 2. Otherwise

intervention a perfect signal that the monitor sent message $m = 1$. Hence, optimal policy always has a baseline level of intervention. The fact that σ_1 is strictly below 1 at the optimum is more surprising. This allows to minimize the cost of baseline interventions σ_0 while keeping the likelihood ratio of intervention rates low.

5.2 Case 2: the righteous monitor

We now turn to the case where

$$\forall m \in \{0,1\}, \quad \mathbb{E}(v_M|c=0,m) \leq 0 \quad \text{and} \quad \mathbb{E}(v_M|c=1,m) \geq 0.$$

In words, the monitor dislikes intervention on non-corrupt agents, and values intervention on corrupt agents regardless of the messages she sends. This corresponds to environments in which the monitor has intrinsic preferences for eliminating corruption.

The first preliminary results are identical to the case of the fearful monitor.

Lemma 9. *Any optimal intervention profile that induces $c = 0$, exhibits $\sigma_1 \geq \sigma_0$. At the optimum, there is no retaliation on the equilibrium path and $m^* = c^*$.*

Conditional on non-corruption, the principal's objective can be written as

$$\max_{\sigma} \sigma_0 \mathbb{E}(v_P|c=0,m=0)$$
$$s.t. \;\; c^*(\sigma) = 0.$$

The next lemma identifies a significant difference with the cases treated in Sections 3 and 5.1.

Lemma 10. *There is no cost of inducing truth-telling: for all σ such that $\sigma_1 \geq \sigma_0$, $K_{c,m=c} = 0$. The cost $K_{1,0}(\lambda)$ of inducing message "non-corrupt" for a corrupt agent need not be decreasing in λ.*

The first point follows from the fact that conditional on intervention, the monitor has incentives to be truthful and from the fact that truthfulness maximizes the probability of intervention. The second part indicates that now it might be the case that increasing the likelihood ratio λ can actually increase the costs of manipulating information, in a clear contrast to the case of the fearful monitor. Indeed, recall that the incentive compatibility constraint of the monitor can be expressed as

$$\lambda \left[\mathbb{E}(v_M|c=1,m=1) - \mathbb{E}(r|c=1,m=1)\right] - \left[\mathbb{E}(v_M|c=1,m=0) - \mathbb{E}(r|c=1,m=0)\right] \leq 0.$$

In contrast to the case of the fearful monitor, this constraint can now bind with $\mathbb{E}(v_M|c=1,m=0) - \mathbb{E}(r|c=1,m=0) > 0$. For instance, this will be the case when $f(z_A|c=1,m=0)$ and $f(z_A|c=1,m=1)$ are sufficiently informative of the monitor's message to concentrate retaliation on the event where $m=1$ and set $\mathbb{E}(r|c=1,m=0)$ close to 0. In this case, when the constraint binds, we have that $\lambda \left[\mathbb{E}(v_M|c=1,m=1) - \mathbb{E}(r|c=1,m=1)\right] > 0$ and an increase in λ increases the cost of silencing the monitor.

More intuitively, an increase in λ means that under message $m=1$ the monitor is relatively more likely to experience both intervention and retaliation. In Section 5.1, both intervention and retaliation are costly to the monitor and increasing λ has an unambiguous effect on incentives. In the current case, intervention is valuable to the monitor and the effect of increasing λ becomes ambiguous. As a consequence, while there is a global trade-off between using and eliciting information (this follows from Lemma 2) this trade-off may sometimes fail to hold locally.

Let us now examine the consequences of this potential failure on optimal intervention strategies and behavior.

Truthtelling. Given a profile σ above the 45^o line, a non-corrupt agent will choose to induce message $m=0$. Given corruption decision $c=1$, the agent will manipulate messages

if and only if

$$\begin{aligned} &\sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda)] \geq \sigma_1 \mathbb{E}(v_A|c=1) \\ \iff &\mathbb{E}(v_A|c=1)(\lambda - 1) + K_{1,0}(\lambda) \leq 0. \end{aligned} \tag{7}$$

The fact that $K_{1,0}(\lambda)$ need not be decreasing in λ, implies that now there may be multiple disjoint intervals of values λ such that (7) holds. Still we know that at the limit where $\lambda \to +\infty$, the agent always engages in manipulation.

Revealed Corruption. When (7) does not hold, the agent will find it optimal to be corrupt if and only if

$$\pi_A + \sigma_1 \mathbb{E}(v_A|c=1) \geq \sigma_0 \mathbb{E}(v_A|c=0) \iff \sigma_1 \leq \frac{\mathbb{E}(v_A|c=0)}{\mathbb{E}(v_A|c=1)}\sigma_0 + \frac{\pi_A}{-\mathbb{E}(v_A|c=1)}. \tag{8}$$

Condition (8) is identical to condition (5) obtained in Section 5.1. It's boundary is a line with slope and intercept within $(0, 1)$.

Unreported Corruption. In contrast, when (7) holds, the agent will be corrupt and manipulate messages if and only if

$$\begin{aligned} &\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda)] \geq \sigma_0 \mathbb{E}(v_A|c=0) \\ \iff &\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - \mathbb{E}(v_A|c=0) - K_{1,0}(\lambda)] \geq 0. \end{aligned} \tag{9}$$

Since $K_{1,0}(\lambda)$ is not necessarily decreasing anymore, the boundary of the set described by (9) need not be a function. However, we know from Theorem 1 that this curve can only cross each ray from the origin once.

Optimum. In the current setting, increasing λ may increase the cost of manipulation and therefore deter corruption. Denote by λ_T the highest value λ such that (7) holds. If increasing λ above λ_T does not increase the cost of manipulation the optimal intervention profile will be such that (7) and (8) hold with equality. Otherwise the optimum may be attained for a value of σ such that $\lambda > \lambda_T$ and there will be no reported corruption in a neighborhood of σ.

Lemma 11. *(i) Define λ_T by $\max\{\lambda|(\lambda-1)\mathbb{E}(v_A|c=1)+K_{1,0}(\lambda)=0\}$. If $\max_{\lambda\geq\lambda_T} K_{1,0}(\lambda) \leq K_{1,0}(\lambda_T)$, then the optimum intervention profile σ that induces non-corruption is characterized by*

$$\pi_A + \sigma_0[\lambda_T\mathbb{E}(v_A|c=1) - \mathbb{E}(v_A|c=0)] = 0 \tag{10}$$

$$\sigma_1 = \lambda_T\sigma_0 \tag{11}$$

and for all $\epsilon > 0$ and all $m \in \{0,1\}$, there exists σ' such that $||\sigma - \sigma'|| < \epsilon$ and $m^*(\sigma') = m$.

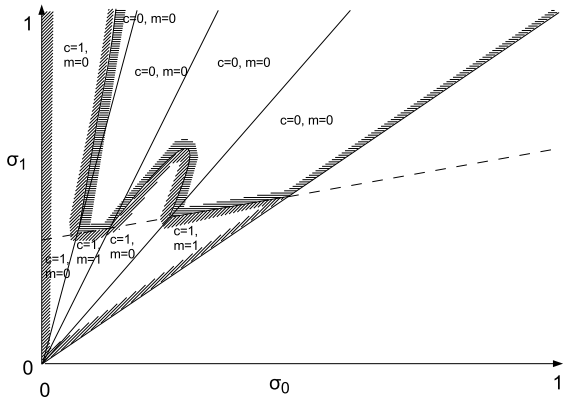
(ii) Otherwise, there exists $\epsilon > 0$ such that for all σ' satisfying $||\sigma - \sigma'|| < \epsilon$, $m^*(\sigma') = m^*(\sigma)$.

The different configurations highlighted in Lemma 11 are illustrated in Figure 3. Note that in this setting message data is not sufficient to identify the optimum intervention profile.

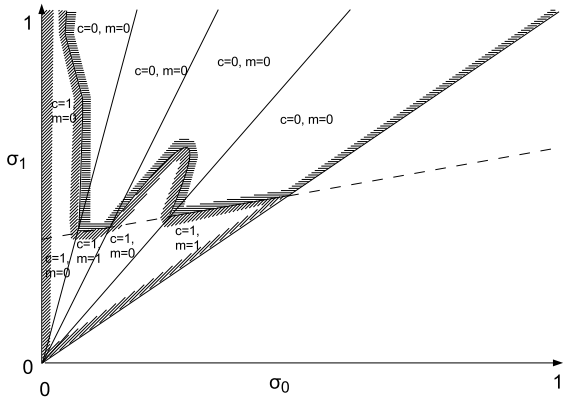
5.3 Case 3: the covetous monitor

The final case we examine corresponds to

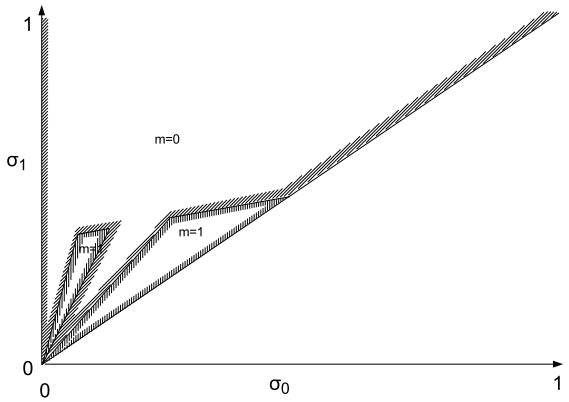
$$\mathbb{E}(v_M|c=0,m) \geq 0.$$



(a) Behavior and messages—identifiable optimum.



(b) Behavior and messages—non-identifiable optimum.



(c) Messages only.

Figure 3: righteous monitor

In other terms, the monitor values intervention on non-corrupt agents (preferences for intervention over corrupt agents may be arbitrary). This may be the case if the monitor hopes to obtain the agent's position if the latter is discredited in a corruption scandal. Alternatively, the monitor may be part of a corruption network and wants to punish honest bureaucrats.

This setting features an obvious difficulty. The monitor might now want to send $m = 1$ when the agent is non-corrupt. Hence, non-corrupt agents may now need to engage in retaliation against the monitor upon intervention, in order to force the monitor to be truthful. In previous cases, analysis was simplified by the fact that there was no retaliation in equilibrium. Unfortunately this is no longer the case. Still the structure described in Theorem 1 survives. In addition, the following result holds.

Lemma 12. *There exists $\overline{\lambda}$ such that for all $\lambda > \overline{\lambda}$, $m = 0$ and the agent is corrupt. There exists $\underline{\lambda} > 1$ such that for all $\lambda \in (1, \underline{\lambda})$, the monitor truthfully reveals the agent's action.*

In words, as in previous cases, when the likelihood ratio of intervention rates is sufficiently high, the agent will be corrupt and induce the message corresponding to the least intervention probability. In turn, when the likelihood ratio of intervention rates is sufficiently close to 1, the monitor will truthfully report the agent's behavior.

6 Identification and Robust Intervention

The stable structure of equilibrium messages suggests that there may be robust policy recommendations for a principal that does not have full knowledge of the underlying parameters of the problem. Let $E = (v_A, v_M, k_A, f(\cdot))$ denote the underlying payoff environment and consider a principal who ignores E. The restrictions that she can impose are those corresponding to the message data she receives for different σ which corresponds to her minimum guaranteed feedback. This would for instance be the case if the consequences z_P take time to realize.

Let $\mathcal{E}$ denote the set of possible environments E. Define $S = \{\sigma \in [0,1]^2 \text{ s.t. } \sigma_1 \geq \sigma_0\}$. For any environment E, let $D_E = \{(\sigma, m_E^*(\sigma))|\sigma \in S\}$ denote messaging data in that environment, and $\mathcal{E}|D = \{E|D_E = D\}$ the set of environments E consistent with messaging data D.

Theorem 2. *Given data D consider $\sigma^\star$ solving*

$$\min \sigma_0 \left| \sigma \in S, \text{ and} \begin{cases} \forall \rho \in [0,1),\ m^*(\rho\sigma) \neq m^*(\sigma) \\ \forall \rho \in [1,+\infty),\ m^*(\rho\sigma) = m^*(\sigma) \end{cases} \right. .$$

Intervention profile $\sigma^\star$ also solves

$$\max_{\sigma \in S} \min_{E \in \mathcal{E}|D} \mathbb{E}(u_P|c=0).$$

The intuition for this theorem is best built by looking at Figures 2 and 3. If the only feedback available is messages, the following is a robust search strategy. Find the lowest value σ_0 consistent with a value σ_1 such that messages change around (σ_0, σ_1). Such an intervention profile is guaranteed to suppress corruption. However, note also that the optimum need not be attainable using only data from messages. This is the case represented in Figure 3(b): the optimal intervention profile is in a portion of the space where messages are locally constant.

7 Conclusion

We model the problem of a principal who hopes to exploit messages from informed monitors to target intervention against a potentially corrupt agent. The difficulty is that the agent can dissuade the monitor from informing the principal by threatening to retaliate conditional on intervention.

In this setting, intervention becomes a signal of the monitor's behavior which the agent can use to improve his own incentive provision. As a consequence, optimal intervention

strategies garbles information in two ways. First, there will necessarily be a strictly positive baseline rate of auditing following the message "non-corrupt". Second, the principal need not intervene with probability one following the message "corrupt". We also show that it is possible to construct tests to rule out unreported corruption using only message data.

In this model we leave out a number of issues in order to highlight the interplay between intervention, retaliation, and information provision. First, we assume that players are homogeneous. Having agents with different propensities of corruption would make many of our results less stark. For instance there may be retaliation on the equilibrium path. Second, we do not model potential collusion either with the monitor, or with possible external auditors in charge of implementing the intervention. We hope to make progress on these questions in future work.

A Proofs

Proof of Lemma 1: Indeed, the agent's payoff from being corrupt is $\pi_A + \mathbb{E}[v_A|c=1] \leq 0$ whereas the agent's payoff from non-corruption is 0. Retaliation is costly and carries no valued added to the agent. There is no intervention on the equilibrium path, so that the principal achieves its highest possible payoff. ■

Proof of Lemma 2: Assume that $\sigma_0 = 0$. Consider a corrupt agent with constant retaliation strategy $r(z_A) = \overline{r}$ whenever there is intervention. The monitor sends message $m = 0$ if and only if

$$\begin{aligned} & \sigma_1[\mathbb{E}(v_M|c=1,m=1) - \overline{r}) \leq \sigma_0[\mathbb{E}(v_M|c=1,m=0) - \overline{r}] \\ \iff & \sigma_1\mathbb{E}(v_M|c=1,m=1) - \sigma_0\mathbb{E}(v_M|c=1,m=0) \leq (\sigma_1 - \sigma_0)\overline{r}, \end{aligned}$$

which will hold for $\overline{r}$ sufficiently large since $\sigma_1 > \sigma_0$. For such a value $\overline{r}$, the monitor sends

message $m = 0$ and the intervention and retaliation costs on the equilibrium path are equal to 0. Hence the agent can guarantee its maximum payoff π_A. An identical proof holds if $\sigma_1 = 0$ and $\sigma_0 > 0$. ■

Proof of Lemma 3: Indeed, the principal's highest possible payoff conditional on corruption is π_P which is ensured by not intervening. If $\sigma_m = 0$, the agent can ensure at message m is sent with probability 1 at no cost, by lemma 2. In this case there is neither retaliation nor intervention. If $\sigma_1 = \sigma_0 = 0$ the implications are immediate. ■

Proof of Lemma 4: We begin with point (i). If message m is costly to induce, then message $\neg m$ can be induced without threat of retaliation. If in addition $\sigma_m \geq \sigma_{\neg m}$ then inducing message $\neg m$ will dominate inducing message m. Turn to point (ii). Assume that the agent exerts some retaliation to induce message m. It must be that,

$$\Delta \equiv -(1-\sigma_m)\mathbb{E}(r|i=0) + \sigma_m(\mathbb{E}(v_M|c,m) - \mathbb{E}(r|c,m)) + (1-\sigma_{\neg m})\mathbb{E}(r|i=0)$$
$$- \sigma_{\neg m}(\mathbb{E}(v_M|c,\neg m) - \mathbb{E}(r|c,\neg m)) \geq 0.$$

By point (i) since the agent is using costly retaliation, it must be that $\sigma_m < \sigma_{\neg m}$. Hence Δ is decreasing in $\mathbb{E}(r|i=0)$ and setting $\mathbb{E}(r|i=0) = 0$ keeps the monitor's message constant, while saving on retaliation costs. Therefore the optimal retaliation scheme inducing message m does not exhibit retaliation conditional on no intervention. ■

Proof of Lemma 5: The result is immediate. ■

Proof of Theorem 1: We begin with point (i). Assume that $m = 1$. Conditional on c,

the set of intervention profiles σ such that $m^*(c,\sigma)=1$ is the set of profiles σ such that

$$\begin{aligned}&\sigma_1\left[\mathbb{E}(v_A|c)-K_{c,1}(\lambda)\right]\geq\sigma_0\left[\mathbb{E}(v_A|c)-K_{c,0}(\lambda)\right]\\ \iff&(\lambda-1)\mathbb{E}(v_A|c)-K_{c,1}(\lambda)+K_{c,0}(\lambda)\geq 0.\end{aligned}\tag{12}$$

Hence profile σ induces message $m*(c,\sigma)=1$ if and only if ratio $\lambda=\sigma_1/\sigma_0$ satisfies (12), which characterizes a cone. An identical proof holds for $m=0$.

We now turn to point (ii). We prove that $\forall\rho\in[1,+\infty)$, $c^*(\sigma)=0\Rightarrow c^*(\rho\sigma)=0$, which necessarily implies that $\forall\rho\in[0,1]$, $c^*(\sigma)=1\Rightarrow c^*(\rho\sigma)=1$. Consider σ such that $c^*(\sigma)=0$. It must be that

$$\pi_A+\max_{m\in\{0,1\}}\sigma_m(\mathbb{E}(v_A|1)-K_{1,m}(\lambda))\leq\max_{m\in\{0,1\}}\sigma_m(\mathbb{E}(v_A|0)-K_{0,m}(\lambda))\iff\quad\pi_A+\sigma_0\Psi(\lambda)\leq 0$$

where

$$\begin{aligned}\Psi(\lambda)\equiv&\max\left\{\mathbb{E}(v_A|1)-K_{1,0}(\lambda),\lambda(\mathbb{E}(v_A|1)-K_{1,1}(\lambda))\right\}\\&-\max\left\{\mathbb{E}(v_A|0)-K_{0,0}(\lambda),\lambda(\mathbb{E}(v_A|0)-K_{0,1}(\lambda))\right\}.\end{aligned}$$

Since $\pi_A>0$, it follows from $\pi_A+\sigma_0\Psi(\lambda)\leq 0$ that $\Psi(\lambda)<0$, hence for any value $\rho\geq 1$, $\pi_A+\rho\sigma_0\Psi(\lambda)\leq 0$, and therefore it must be that $c^*(\rho\sigma)=0$. ■

Proof of Corollary 1: Indeed, by Theorem 1(ii) we have that if $c^*(\sigma)=1$, then for all $\rho\leq 1$, $c^*(\rho\sigma)=1$, and hence by Theorem 1(i), we have that $m^*(\sigma)=m^*(\rho\sigma)$. An additional testable fact is that if $c^*(\sigma)=0$, for all $\rho>1$, $m^*(\sigma)=m^*(\rho\sigma)$. The proof is identical. ■

Proof of Lemma 6: We begin by showing that $\sigma_1\geq\sigma_0$. Assume otherwise. If $\sigma_1<\sigma_0$,

since

$$\sigma_1(\mathbb{E}(v_M|c=1,m=1)-r)-\sigma_0(\mathbb{E}(v_M|c=1,m=0,)-r)$$

is increasing in r, and it is positive when $r=0$, the agent cannot induce $m=0$ if $c=1$. Hence, $K_{1,1}(\lambda)=0$. Hence, if $c=1$, the payoff of the agent is $\pi_A+\sigma_1\mathbb{E}(v_A|c=1)$. For corruption not to be optimal, it must be that

$$\pi_A+\sigma_1\mathbb{E}(v_A|1)\leq \max\{\sigma_1(\mathbb{E}(v_A|0)-K_{0,1}(\lambda)),\sigma_0(\mathbb{E}(v_A|0)-K_{0,0}(\lambda))\}.$$

But note that

$$\pi_A+\sigma_1\mathbb{E}(v_A|1)\leq \sigma_1(\mathbb{E}(v_A|0)-K_{0,1}(\lambda))$$

is impossible since $\pi_A>0$ and $K_{0,1}(\lambda)\geq 0$. Hence it must be the case that

$$\pi_A+\sigma_1\mathbb{E}(v_A|1)\leq \sigma_0(\mathbb{E}(v_A|0)-K_{0,0}(\lambda)).$$

But note that in this case we can reduce σ_0 until we reach $\sigma_0=\sigma_1$. This leaves the right hand side untouched but increases the left hand side since $K_{0,0}(1)=0$. This modification must be better for the principal since it substantially reduces intervention.

Let us now show that at the optimum, there is no retaliation on the equilibrium path. If in equilibrium the agent is corrupt, it follows from Lemma 3 that there is no retaliation on the equilibrium path. Assume now that in equilibrium the agent is not corrupt. The non-corrupt agent induces the message m that maximizes $\sigma_m(\mathbb{E}(v_A|0)-K_{0,m}(\lambda))$. Since $\lambda\geq 1$ and $\mathbb{E}(v_M|0,m)\leq 0$, we have

$$\sigma_1\mathbb{E}(v_M|0,1)\leq \sigma_0\mathbb{E}(v_M|0,0),$$

and therefore $K_{0,0}(\lambda)=0$. It follows that the non-corrupt agent chooses to induce message 0 and can do so without resorting to costly retaliation. ■

Proof of Lemma 7: Cost $K_{1,0}(\lambda)$ can be expressed as

$$K_{1,0}(\lambda) = \inf_{r:Z_A \to [0,+\infty)} \int_{Z_A} k_A(r(z_A)) f(z_A|c,m) \mathrm{d}z_A$$

s.t.

$$\lambda \left[\mathbb{E}(v_M|c=1,m=1) - \mathbb{E}(r|c=1,m=1)\right] - \left[\mathbb{E}(v_M|c=1,m=0) - \mathbb{E}(r|c=1,m=0)\right] \leq 0. \tag{13}$$

Since we have $\mathbb{E}(v_M|c=1,m=1) - \mathbb{E}(r|c=1,m=1) \leq 0$, it follows that if retaliation profile r satisfies (13) for $\lambda \geq 0$, then it satisfies (13) for all $\lambda' \geq \lambda$. Hence $K_{1,0}(\lambda)$ is necessarily decreasing in λ. If $\lambda = 1$, then (13) cannot be satisfied with $r = 0$ since by assumption, $\mathbb{E}(v_M|c=1,m=1) > \mathbb{E}(v_M|c=1,m=0)$. ■

Proof of Lemma 8: This follows from the fact that the boundary of $\sigma_1(\sigma_0)$ of the manipulative corruption region is increasing in σ_0. ■

Proof of Corollary 2: This optimal intervention profile is at the crossing of 4 and 5. These are both increasing straight lines. 4 is a ray that starts in the origin and has slope larger than one since $K_{1,0}(1) > 0$. This slope also has to be smaller than infinity because of lemma 2. As noted, 5 is increasing with an intercept and slope both in $(0,1)$. As sufficient condition for the crossing to be interior is that 5 is below 1 at $\sigma_0 = 1$. But this is ensured by assumption 1. ■

Proof of Lemma 9: The proof is identical to that of lemma 6. ■

Proof of Lemma 10: We begin by showing that whenever $\sigma_1 \geq \sigma_0$, $K_{c,m=c}(\lambda) = 0$ for

any corruption decision c. Indeed, we have that

$$\sigma_1 \mathbb{E}(v_M|c=1,m=1) \geq \sigma_0 \mathbb{E}(v_M|c=1,m=0)$$
$$\sigma_1 \mathbb{E}(v_M|c=0,m=1) \leq \sigma_0 \mathbb{E}(v_M|c=0,m=0)$$

which implies that in the absence of retaliation, the monitor will choose to report the truth. It follows that $K_{c,m=c}(\lambda) = 0$ for any corruption decision c.

Cost $K_{1,0}(\lambda)$ can be expressed as

$$K_{1,0}(\lambda) = \inf_{r:Z_A \to [0,+\infty)} \int_{Z_A} k_A(r(z_A)) f(z_A|c,m) \mathrm{d}z_A$$

s.t.

$$\lambda \left[\mathbb{E}(v_M|c=1,m=1) - \mathbb{E}(r|c=1,m=1)\right] - \left[\mathbb{E}(v_M|c=1,m=0) - \mathbb{E}(r|c=1,m=0)\right] \leq 0. \tag{14}$$

Now notice that it can be the case that (14) is binding and $\mathbb{E}(v_M|c=1,m=0) - \mathbb{E}(r|c=1,m=0)$ is positive. This would mean that $\lambda \left[\mathbb{E}(v_M|c=1,m=1) - \mathbb{E}(r|c=1,m=1)\right]$ is also positive. In such cases, an increase in λ would break the constraint and the agent would need to increase $r(z_A)$ for some z_A thus increasing $K_{1,0}(\lambda)$. ■

Proof of Lemma 11: Intervention profiles that dissuade corruption either dissuade truthful or manipulative corruption:

truthful corruption	$(\lambda - 1)\mathbb{E}(v_A|c=1) + K_{1,0}(\lambda) \geq 0$	(15)
	$\pi_A + \sigma_0[\lambda \mathbb{E}(v_A|c=1) - \mathbb{E}(v_A|c=0)] \leq 0$	(16)
manipulative corruption	$(\lambda - 1)\mathbb{E}(v_A|c=1) + K_{1,0}(\lambda) \leq 0$	(17)
	$\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda) - \mathbb{E}(v_A|c=0)] \leq 0.$	(18)

Since $\mathbb{E}(v_A|c=1) \leq 0$, the optimal intervention profile among profiles corresponding to truthful corruption corresponds to the highest value λ_T such that $(\lambda-1)\mathbb{E}(v_A|c=1) + K_{1,0}(\lambda) \geq 0$, which implies that $(\lambda_T - 1)\mathbb{E}(v_A|c=1) + K_{1,0}(\lambda_T) = 0$. This implies that

$$\pi_A + \sigma_0[\lambda_T \mathbb{E}(v_A|c=1) - \mathbb{E}(v_A|c=0)] = \pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda_T) - \mathbb{E}(v_A|c=0)].$$

Any $\lambda < \lambda_T$ such that manipulation is optimal conditional on corruption is such that $(\lambda_1)\mathbb{E}(v_A|c=1)+K_{1,0}(\lambda) \leq 0$. Taking differences, we obtain that $K_{1,0}(\lambda_T) \geq K_{1,0}(\lambda)-(\lambda_T-\lambda)\mathbb{E}(v_A|c=1) \geq K_{1,0}(\lambda)$. Hence it follows that $\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda) - \mathbb{E}(v_A|c=0)] \leq 0$ implies $\pi_A + \sigma_0[\mathbb{E}(v_A|c=1) - K_{1,0}(\lambda_T) - \mathbb{E}(v_A|c=0)] \leq 0$. Therefore λ_T sustains no-corruption at a lower intervention rate σ_0 than λ.

Any $\lambda > \lambda_T$ is such that $(\lambda-1)\mathbb{E}(v_A|c=1) + K_{1,0}(\lambda) \leq 0$ by construction. Ratio λ sustains no corruption at a lower intervention rate σ_0 than λ_T if and only if, $K_{1,0}(\lambda) > K_{1,0}(\lambda_T)$. This proves points (i) and (ii), with the messaging variability results being immediate from inspection. ■

Proof of Lemma 12: The first result follows from the fact that for λ large enough the cost of inducing message $m=0$ goes to zero while the benefits are bounded away from 0. The second result follows from the fact that for λ close enough to 1, the optimal retaliation policy is to set $r=0$, furthermore, for λ sufficiently close to 1, the monitor has incentives for truth-telling: $\mathbb{E}(v_M|c, m=c) > \mathbb{E}(v_M|c, m=\neg c) \Rightarrow \sigma_m \mathbb{E}(v_M|c, m=c) > \sigma_{\neg m}\mathbb{E}(v_M|c, m=\neg c)$. ■

References

Acemoglu, D. and T. Verdier (1998): "Property rights, Corruption and the Allocation of Talent: a general equilibrium approach," *Economic Journal*, 108, 1381–1403.

——— (2000): "The Choice between Market Failures and Corruption," *American Economic Review*, 194–211.

Banerjee, A. and E. Duflo (2006): "Addressing Absence," *Journal of Economic Perspectives*, 20, 117–132.

Duflo, E., R. Hanna, and S. Ryan (forthcoming): "Incentives work: Getting teachers to come to school," *American Economic Review*.

Eeckhout, J., N. Persico, and P. Todd (2010): "A theory of optimal random crackdowns," *The American Economic Review*, 100, 1104–1135.

Ensminger, J. (2012): "Inside Corruption Networks: Community Driven Development in the Village," *Unpublished manuscript*.

Ghosh, A. and A. Roth (2010): "Selling privacy at auction," *Arxiv preprint arXiv:1011.1375*.

Gradwohl, R. (2012): "Privacy in Implementation," .

Izmalkov, S., M. Lepinski, and S. Micali (2011): "Perfect implementation," *Games and Economic Behavior*, 71, 121–140.

Laffont, J. and D. Martimort (1997): "Collusion under asymmetric information," *Econometrica: Journal of the Econometric Society*, 875–911.

Mauro, P. (1995): "Corruption and Growth," *Quartely Journal of Economics*, 110, 681–712.

Nissim, K., C. Orlandi, and R. Smorodinsky (2011): "Privacy-aware mechanism design," *Arxiv preprint arXiv:1111.3350*.

Olken, B. (2007): "Monitoring corruption: evidence from a field experiment in Indonesia," *Journal of Political Economy*, 115, 200–249.

Olken, B. and R. Pande (2011): "Corruption in Developing Countries," .

Punch, M. (2009): *Police Corruption: Deviance, Accountability and Reform in Policing*, Willan Publishing.

Rahman, D. (forthcoming): "But who will monitor the monitor?" *The American Economic Review*.

Shleifer, A. and R. W. Vishny (1993): "Corruption," *Quartely Journal of Economics*, 108, 599–617.

Tirole, J. (1986): "Hierarchies and bureaucracies: On the role of collusion in organizations," *JL Econ. & Org.*, 2, 181.