Corruption, Intimidation and Whistleblowing: A Theory of Inference from Unverifiable Reports*

Sylvain Chassang — Princeton University

Gerard Padró i Miquel† — London School of Economics

August 27, 2013.

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. As a result, intervention policies that are very responsive to the monitor's message can give rise to silent corruption in which the agent dissuades informative reporting. Successful intervention policies must therefore garble the information provided by monitors. We show that even if hard evidence is unavailable and monitors have heterogeneous incentives to (mis)report, it is possible to establish robust bounds on equilibrium corruption using only non-verifiable reports. Our analysis suggests a simple heuristic to calibrate intervention policies: first get monitors to complain, then scale up enforcement while keeping the information content of intervention constant.

KEYWORDS: corruption, whistleblowing, plausible deniability, inference, prior-free policy design.

*We are indebted to Michael Callen, Hans Christensen, Ray Fisman, Matt Gentzkow, Navin Kartik, Jesse Shapiro, as well as seminar audiences at Columbia, the Institute for Advanced Study, the Nemmers Prize Conference, NYU, ThReD, and the UCSD workshop on Cellular Technology, Security and Governance for helpful conversations. Chassang gratefully acknowledges the hospitality of the University of Chicago Booth School of Business, as well as support from the Alfred P. Sloan Foundation and the National Science Foundation under grant SES-1156154. Padró i Miquel acknowledges financial support from the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Starting Grant Agreement no. 283837.

†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.[1] 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 effective retaliation by corrupt agents and limits endogenous information provision. As a consequence there is a trade-off between eliciting information and using that information efficiently. 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 the relationship between intervention, corruption and whistleblowing, and suggest ways to identify effective intervention strategies using only unverifiable reports.

Our framework encompasses various forms of corruption such as 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; whistleblowers may be harrassed, see their careers derailed, or get sued for defamation; police officers suspected of collaborating with Internal Affairs may have their life threatened by lack of prompt support.[2] 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 following complaints

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

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

to authorities.[3] In addition, monitors may fear that anonymity is not properly ensured and that untrustworthy institutions may leak the source of complaints to the agent or one of his associates. All these situations exhibit two features that are key to our analysis: (1) there is significant information about corrupt agents which the principal wants to obtain; (2) the individuals who have this information and are able to pass it on to the principal can be punished by the agent.

Our model considers a dynamic game played by a principal, an agent and a 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 subsequent observables — including whether or not he is the subject of an investigation — and makes a corruption decision. The monitor observes the corruption behavior of the agent and chooses what message to send to the principal. Finally, intervention and retaliation are realized according to the commitments of both the principal and the agent.

A key element of our modeling 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 is determined by an exogenous judiciary process. Similarly, whistleblower protection schemes may not fully shield the monitor against indirect punishments such as ostracism, or harassment, and supposedly anonymous information may be leaked. Furthermore, we do not assume that the monitor necessarily desires intervention against corrupt agents. For instance, in a development context, corruption scandals may lead to a withdrawal of funding altogether, which hurts citizens even if they were getting only a small share of funds. We also allow for

[3]For instance, Ensminger (2013) suggests that egregious corruption affecting the World Bank's arid land program were not reported by the local Kenyan communities that suffered from it for fear of being cut off from subsequent projects.

the possibility of covetous monitors, i.e. monitors who benefit from having an honest agent investigated. Whether they hold a personal grudge, or seek to discredit a competitor, such covetous monitors may report that corruption is occurring even when it is not the case.

Our analysis emphasizes two sets of results. The first is that any effective intervention strategy must garble the information provided by the monitor. Indeed, because the principal's behavior is correlated with the monitor's message, it is a signal that the agent can exploit to resolve his own agency problem vis-à-vis the monitor: when the likelihood ratio of intervention rates under messages "corrupt" and "not corrupt" is high, the threat of retaliation conditional on intervention dissuades the monitor to send informative messages at little equilibrium cost to the agent. An immediate consequence is that the likelihood of intervention against non-corrupt agents must be bounded away from zero. In addition, it may be optimal not to intervene against agents reported as corrupt, since this allows to reduce costly intervention on non-corrupt agents while keeping the information content of intervention low.

Our main set of results characterizes the geometry of reporting and corruption decisions as a function of intervention rates, in a rich environment with both heterogeneous payoffs and heterogeneous beliefs. We show that the region of the intervention-strategy space in which corruption occurs is star-shaped around the origin, and that keeping corruption behavior constant, messages by agents depend only on the likelihood ratio of intervention rates. We show how to exploit these properties to obtain sharp bounds on corruption using non-verifiable reports alone. This analysis suggests that a useful rule-of-thumb to determine appropriate intervention policies is to first provide sufficient plausible deniability that monitors are willing to complain, and then scale up enforcement while keeping the information content of intervention constant.

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 issues affecting public service provision and public expenditure in education or health (see Olken and Pande (2011) for a recent review), while recent experimental evidence suggests that appropriate incentive design can reduce misbehavior (Olken (2007), Duflo et al. (forthcoming)). In our view, one key aspect of corruption is that even when there is strong suspicion that it is occurring, there seems to be little direct and actionable evidence flowing back to the relevant principals (see for instance Ensminger (2013) who emphasizes the role of threats and failed information channels in recent corruption scandals affecting community driven development projects).[5] We show that correct policy design is essential to keep information channels open under the threat of retaliation, and we suggest ways to measure underlying corruption using only unverifiable messages.

Our work is closely related to that of Tirole (1986), Laffont and Martimort (1997) or Prendergast (2000) on principal-agent problems with side-contracting between the agent and the monitor. Our approach differs in several ways: first, we focus on retaliation, rather than side payments, as the main side-contracting instrument[6]; second we endogenize the difficulty of the side-contracting problem between the agent and the monitor; third, we allow for non-verifiable messages and monitors with heterogeneous motives; fourth, we focus on inference and seek to establish bounds on unobserved corruption, rather than solve for optimal contracts in specific environments. Our work is also related to that of Rahman

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

[5]In a discussion of why citizens fail to complain about poor public service, Banerjee and Duflo (2006) suggest that "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."

[6]This assumption can be endogenized using the fact that payments are costly on the equilibrium path. It plays an important role in the analysis as discussed in Appendix A.

(2012) who also considers agency problems with non-verifiable reports, and emphasizes the value of random recommendation-based incentives to jointly incentivize effort provision by the agent and by the monitor supposed to evaluate the agent. However, Rahman (2012) excludes the possibility of side contracting between the agent and the monitor. As a result, 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.[7] Finally our work shares much of its motivation with the seminal work of Warner (1965) on the role of plausible deniability in survey design, and the recent work of Izmalkov et al. (2011), Ghosh and Roth (2010), Nissim et al. (2011), or Gradwohl (2012) on privacy in mechanism design.

The paper is structured as follows: Section 2 introduces our model and delineates the main points of our analysis using a simple example; Section 3 introduces our general framework which allows for rich incomplete information; Section 4 establishes general patterns of corruption and reporting as the intervention policy varies, and shows how they can be exploited to evaluate unobserved corruption and make policy recommendations; Section 5 discusses potential applications and related implementation challenges. Appendix A presents several extensions. Proofs are contained in Appendix B.

2 An Example

This section introduces our framework and illustrates the mechanics of corruption, intimidation and whistleblowing in the context of a simple example. For clarity, we make several restrictive assumptions, which we generalize in Sections 3 and 4. Specifically, we work under complete information, assume that the monitor is non-covetous (i.e. does not benefit from intervention against an honest agent), and that the agent does not have access to

[7] 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.

side information about the monitor's message, except that provided through the principal's intervention strategy.

2.1 Setup

Players, timing, and actions. There are three players: a principal P, an agent A and a monitor M. The timing of actions is as follows.

1. The agent chooses whether to be corrupt ($c = 1$) or not ($c = 0$). The monitor observes 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 payoff consequences for the principal, the agent and the monitor.

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

Observables and payoffs. The monitor costlessly observes the agent's corruption decision $c \in \{0, 1\}$, and can send a message $m \in \{0, 1\}$ to the otherwise uninformed principal. The agent does not observe the monitor's message m, but observes whether the principal triggers an intervention $i \in \{0, 1\}$.[8]

As a function of $c \in \{0, 1\}, i \in \{0, 1\}$ and $r \geq 0$, realized payoffs u_A, u_P and u_M to the agent, principal and monitor take the form

$$u_M = \pi_M \times c + v_M(c, m) \times i - r$$

$$u_A = \pi_A \times c + v_A(c) \times i - k_A(r)$$

$$u_P = \pi_P \times c + v_P(c) \times i$$

[8]Our general framework allows the agent to observe leaks from the institutional process that can be informative of the message m sent by the monitor.

where π_M, π_A, and π_P capture the expected payoff consequences of corruption, v_M, v_A, and v_P capture the expected payoff consequences of intervention, r is the level of retaliation imposed by the agent on the monitor, and $k_A(r)$ is the cost of retaliation to the agent. Payoffs conditional on corruption are such that $\pi_A > 0$ and $\pi_P < 0$. The cost of retaliation $k_A(r)$ is strictly increasing in r, with $k_A(0) = 0$. Payoffs are common-knowledge. We make the following assumption.

Assumption 1. *Expected continuation payoffs following intervention ($i = 1$) satisfy*

$$
\begin{aligned}
& v_M(c=0, m=1) < 0 && \textit{(non-covetous monitor)};\\
& \pi_A + v_A(c=1) < v_A(c=0) \leq 0 && \textit{(dissuasive intervention)};\\
\forall c \in \{0,1\}, \quad & v_M(c, m \neq c) \leq v_M(c, m=c) && \textit{(weak preferences for the truth)};\\
\forall c \in \{0,1\}, \quad & v_P(c) \leq 0 \ \ \textit{and} \ \ \pi_P \leq v_P(c=0) && \textit{(bounded cost of intervention).}
\end{aligned}
$$

The assumptions that there are no covetous monitors — i.e. that the monitor gets a negative continuation payoff $v_M(c = 0, m = 1) < 0$ following intervention on an honest agent — and that certain intervention is sufficient to dissuade the agent from being corrupt, are constraining, but are only made for simplicity. We relax them in the general analysis of Sections 3 and 4. Our two other assumptions are more innocuous. First, we assume that taking intervention as given, the monitor is weakly better off telling the truth. This assumption, which typically comes for free when payoffs are derived from a full mechanism design problem, serves to give an operational meaning to messages $m \in \{0, 1\}$. Second, we assume that intervention is costly to the principal.

Strategies and commitment. Both the principal and the agent can commit to strategies ex ante. Though we do not provide explicit micro-foundations, we think of this commitment power as arising from repeated interaction. The principal is the first mover and commits to an intervention policy $\sigma : m \in \{0, 1\} \mapsto \sigma_m \in [0, 1]$, where $\sigma_m \equiv \text{prob}(i = 1|m)$ is

the likelihood of intervention given message m.[9] Without loss of generality, we focus on strategies such that $\sigma_1 \geq \sigma_0$.

Knowing the principal's intervention strategy σ, the agent takes a corruption decision $c \in \{0,1\}$ and commits to a retaliation policy $r : i \in \{0,1\} \mapsto r(i) \in [0,+\infty)$ as a function of whether or not he observes intervention. The monitor moves last and chooses the message $m \in \{0,1\}$ maximizing her payoffs given the commitments of both the principal and the agent.[10]

We are interested in characterizing patterns of corruption and information transmission as the principal's policy σ changes. We also solve for the principal's optimal intervention policy σ and show that it must be interior.

Reduced-form payoffs. It is important to note that while we take payoffs upon intervention as exogenous, this does not mean that our approach is inconsistent with a broader mechanism design problem in which payoffs upon intervention v_A and v_M are also policy variables affected the principal. Indeed, we place few restrictions on reduced-form payoffs, and they can be thought of as being determined in a first optimization stage, before determining intervention patterns σ.

Formally, if $\mathcal{V}$ denotes the set of feasible payoff structures $v = (v_A, v_M)$, Σ the set of possible intervention policies σ, and $c^*(v,\sigma)$ the agent's equilibrium behavior under payoff structure v and policy σ, the principal can be thought of as solving

$$\max_{v\in\mathcal{V},\sigma\in\Sigma} \mathbb{E}[u_P|\sigma, c^*(v,\sigma)] = \max_{v\in\mathcal{V}} \max_{\sigma\in\Sigma} \mathbb{E}[u_P|\sigma, c^*(v,\sigma)].$$

[9] We assume that the principal can commit to using a mixed strategy. Section 5 discusses credible ways for the principal to do so. In particular, we suggest that mixing can be obtained by garbling the messages provided by the monitor directly at the recording stage, before it even reaches the principal.

[10] The order of moves reflects the various parties ability to make more or less public commitments: the principal can make fully public commitments, whereas the agent can only commit vis-à-vis the monitor: public commitments to retaliate would be directly incriminating.

Provided that payoffs in $\mathcal{V}$ satisfy Assumption 1 (or the relaxed assumptions of Section 3), our analysis applies within the broader mechanism design problem.

Our decision to eschew endogenizing payoffs reflects what we perceive as great heterogeneity in the ability of principals to reliably affect the welfare of involved parties. Indeed, even powerful international organizations such as the World Bank need to go through local judiciary systems to target corrupt agents. For this reason, while payoffs are clearly a first order policy instrument when available, we choose to focus on intervention profiles as our main policy dimension of interest. One that we believe is novel, important, and largely available to principals regardless of external institutional constraints they may be subjected to.

2.2 The Trade-off Between Eliciting and Using Information

To frame the analysis it is useful to contrast the effectiveness of intervention policies when messages are exogenously informative, i.e. when the monitor is an automaton with strategy $m(c) = c$, and when messages are endogenous.

Fact 1 (basic trade-off)**.** *(i) If messages are exogenously informative, i.e.* $\mathbf{m}(c) = c$*, setting* $\sigma_0 = 0$ *and* $\sigma_1 = 1$ *is an optimal policy. There is no corruption and no retaliation in equilibrium.*

(ii) If messages are endogenous, theres exists $\overline{\lambda} > 1$ *such that for any intervention policy* σ *satisfying* $\frac{\sigma_1}{\sigma_0} \geq \overline{\lambda}$*,*

- *the agent is corrupt and commits to retaliate conditional on intervention;*
- *the monitor sends message* $m = 0$*.*

Point (i) follows from Assumption 1, which ensures that the agent refrains from corruption if intervention occurs with high enough probability. Since messages are exogenous,

intervention can be fully responsive to the monitor's message: it provides sufficient incentives for the agent to be honest, and avoids costly intervention on the equilibrium path.

Point (ii) shows that this is no longer the case when messages are endogenous. In this case, when the likelihood ratio $\frac{\sigma_1}{\sigma_0}$ is high, intervention itself becomes a very informative signal of which message the monitor sent. This means that the agent can dissuade the monitor to send message $m = 1$ while keeping incentive costs low, simply by threatening the monitor with high levels of retaliation conditional on intervention.

To prevent corruption, the principal must therefore commit to trigger intervention with sufficiently high probability even when she is obtains message $m = 0$. This makes the agent's own incentive problem more difficult to resolve, since retaliation must be carried out with positive probability.

An anecdote. The main takeaway from Fact 1 is that a strictly positive baseline rate of intervention $\sigma_0 > 0$ is needed to ensure that information will flow from the monitor to the principal. Indeed, this provides the monitor with plausible deniability, should her message lead to an intervention, which makes incentive provision by the agent harder.

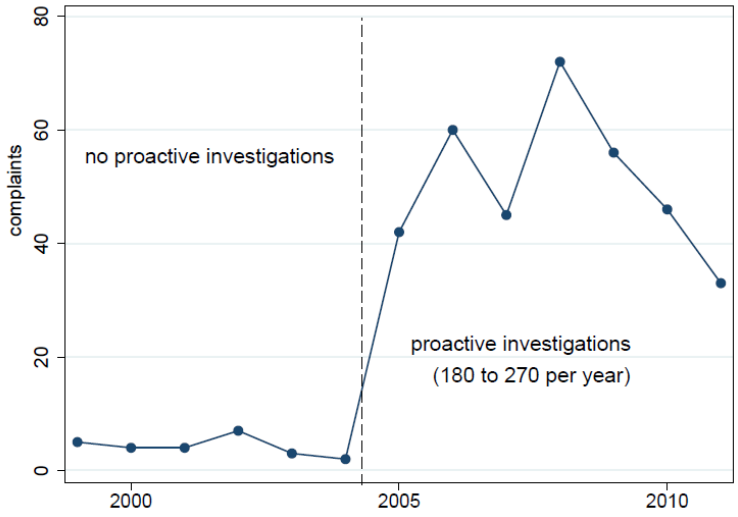
To provide a plausible illustration of how this mechanism may play out in practice, we use the example of recent evolutions in British accounting-oversight policy.[11] We emphasize that the goal here is only to describe the trade-off identified in Fact 1 sufficiently realistically that it can be used to rationalize existing data. This, however, is merely a suggestive anecdote and there are clearly alternative interpretations of the data we discuss.[12]

Between 2004 and 2005 the UK's Financial Reporting Review Panel — the regulatory authority in charge of investing the accounts of publicly owned firms — radically changed its investigation policy. It moved from a purely reactive policy — in which investigations were

[11] We are grateful to Hans Christensen for suggesting this example.

[12] In fact, concurrent changes make this example unsuitable for proper identification. For instance, over a time period covering the data we bring up, accounting standards were being unified across Europe.

only conducted in response to complaints filed by credible agents — to a proactive policy, under which a significant number of firms were investigated each year regardless of whether complaints were filed or not; credible complaints continuing to be investigated as before (Financial Reporting Council, 2004). The change in the number of complaints is large, going from an average of 4 a year in the period from 1999 to 2004, to an average of 50 a year in the period from 2005 to 2011.[13]



This striking pattern can be mapped to our framework as follows. It turns out that the natural monitor of a firm's aggregate accounting behavior is the firm's own auditor. Under a purely reactive system, following intervention, the firm knows that its auditor must have reported it. Of course, this puts the auditor in a difficult position, and is likely to disrupt future business. In contrast, under a proactive system, baseline intervention rates give the auditor plausible deniability should its client be investigated, thereby limiting the damages to long-run cooperation. As a result, proactive investigations allow for higher rates of complaints.

[13]The data is obtained from Brown and Tarca (2007) for years 1999 to 2004, and from the Financial Reporting Review Panel (2005–2011) for years 2005 to 2011.

2.3 Intervention, Reporting and Corruption

We now study in greater details patterns of corruption and information flow as a function of intervention policy σ. Recall that we assumed the monitor was non-covetous, i.e. that $v_M(c=0,m=1)<0$. We proceed by backward induction.

Reporting by the monitor. We begin by clarifying the conditions under which the monitor will report corruption or not. Take as given an intervention profile $\sigma=(\sigma_0,\sigma_1)$, with $\sigma_0<\sigma_1$, and a level of retaliation r conditional on intervention.

We first note that when the agent is not corrupt ($c=0$), it is optimal for the monitor to send message $m=0$ regardless of retaliation level r. Indeed, we necessarily have that

$$\sigma_1[v_M(c=0,m=1)-r]\leq\sigma_0[v_M(c=0,m=0)-r].$$

Note that this relies on the assumption that the monitor is non-covetous. When the monitor may be covetous, even honest agents may threaten to retaliate to ensure that message $m=0$ is sent.

Consider now the case where the agent chooses to be corrupt, i.e. $c=1$. The monitor will report corruption and send message $m=1$ if and only if

$$\sigma_1[v_M(c=1,m=1)-r]\geq\sigma_0[v_M(c=1,m=0)-r].$$

This holds whenever

$$r\leq r_\sigma\equiv\left[\frac{\sigma_1 v_M(c=1,m=1)-\sigma_0 v_M(c=1,m=0)}{\sigma_1-\sigma_0}\right]^+ \tag{1}$$

where by convention $x^+=\max\{x,0\}$. Expression (1) suggests an instructive classification of potential monitors.

If $v_M(c=1,m=1)<0$ the monitor suffers from intervention even against a corrupt agent. As a result, there will be intervention profiles σ such that $r_\sigma = 0$: the monitor prefers to send message $m=0$ even in the absence of retaliation. This possibility is a prominent concern in the context of foreign aid since reports of corruption can cause aid to be withheld (Ensminger, 2013).

If instead $v_M(c=1,m=1)>0$, the monitor values intervention against a corrupt agent, and for any intervention profile σ, positive amounts of retaliation $r_\sigma > 0$ are needed to dissuade the monitor from reporting corruption.

We define $\lambda \equiv \frac{\sigma_1}{\sigma_0}$ and note that r_σ can be expressed only as a function of λ:

$$r_\sigma = r_\lambda \equiv \left[\frac{\lambda v_M(c=1,m=1) - v_M(c=1,m=0)}{\lambda - 1}\right]^+.$$

Note that r_λ is decreasing in likelihood-ratio λ: when the information content of intervention is large, moderate threats of retaliation are sufficient to shut-down reporting.

Information manipulation and corruption. We now examine the agent's behavior. Consider first the agent's incentives to influence reporting conditional on being corrupt, that is, assuming that $c=1$. Since retaliation r is costly to the agent, he either picks $r=0$ and lets the monitor send truthful messages, or picks $r=r_\sigma$ and induces message $m=0$ at the lowest possible cost. Recalling that $\lambda = \frac{\sigma_1}{\sigma_0}$, the agent will manipulate messages through the threat of retaliation if and only if:

$$\begin{aligned} &\sigma_1 v_A(c=1) \leq \sigma_0[v_A(c=1) - k_A(r_\sigma)] \\ \iff\ &\lambda v_A(c=1) \leq v_A(c=1) - k_A(r_\lambda). \end{aligned} \tag{2}$$

Hence, the agent will choose not to be corrupt if and only if

$$\pi_A + \max\{\sigma_1 v_A(c=1), \sigma_0[v_A(c=1) - k_A(r_\sigma)]\} \leq \sigma_0 v_A(c=0). \tag{3}$$

The corresponding patterns of intervention, corruption and reporting are illustrated in Figure 1. For the purpose of inference, we are especially interested in the relationship between reports and underlying corruption. In this example, even though reports are unverifiable and silent corruption is a possibility, variation in reports across different policy choices provides significant information about the underlying amount of corruption.

Consider old and new intervention profiles σ^O and σ^N such that

$$\sigma_0^O < \sigma_0^N, \quad \sigma_1^O < \sigma_1^N, \quad \text{and} \quad \frac{\sigma_1^N}{\sigma_0^N} < \frac{\sigma_1^O}{\sigma_0^O}. \tag{4}$$

We think of these two intervention profiles as policy experiments implemented on different subsamples of a population of agents and monitors.[14] Intervention profile σ^N involves strictly more intervention than σ^O while being less informative about the monitor's report. Let c^O, c^N and m^O, m^N denote the corresponding corruption and reporting decisions in equilibrium (conditional on σ^O and σ^N). The following properties hold.

Fact 2. *(i) There exists $\lambda_0 \geq 1$ such that a corrupt agent induces message $m = 0$ if and only if $\frac{\sigma_1}{\sigma_0} \geq \lambda_0$.*

(ii) If $v_A(c=0) = 0$, then $m^N = 1$ implies that $c^O = 1$.

(iii) For all payoffs satisfying Assumption 1, $\{m^O = 1$ and $m^N = 0\}$ implies that $c^N = 0$.

Point (i) states that corrupt agents shut down information transmission whenever the

[14] Taking seriously this population view of the agency problem, we allow for heterogeneity across agents and monitors in Sections 3 and 4.

likelihood ratio of intervention rates is high enough. Point (ii) has important empirical content. Assume that intervention is costless to non-corrupt agents. If corruption is reported at a new policy σ^N which makes intervention more likely, while also decreasing the likelihood ratio of intervention rates, then it must be that there was corruption (possibly unreported) at the old intervention policy σ^O. This allows us to detect silent corruption. Point (iii) shows how one can identify intervention policies that discourage corruption from unverifiable messages alone: if people complained at the profile σ^O which involved less intervention and less plausible deniability, then a lack of report at policy σ^N can be reliably interpreted as evidence that there is no underlying corruption.

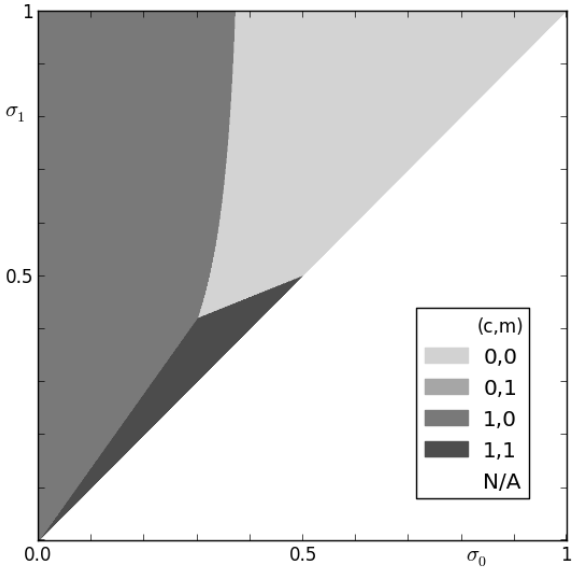


Figure 1: corruption and messages (c, m) as a function of intervention profiles (σ_0, σ_1); payoff specification $\pi_A = 3$, $v_A(c) = -4 - 6c$, $v_M(c, m) = -2 + c(3 + 2m)$, $k_A(r) = 5r$.

Optimal intervention. It is instructive to characterize the optimal intervention profile: we show that the optimal policy involves interior rates of intervention conditional on both messages $m = 0$ and $m = 1$.

Fact 3. *The optimal intervention profile σ^* satisfies* (2) *and* (3) *with equality:*

$$\sigma_1^* = \lambda_0 \sigma_0^* \quad \textit{and} \quad \sigma_1^* = \frac{\pi_A}{-v_A(c=1)} + \sigma_0^* \frac{v_A(c=0)}{v_A(c=1)}.$$

Profile σ^* is interior: $\sigma_0^* \in (0,1)$ and $\sigma_1^* \in (0,1)$. Under policy σ^*, there is no corruption and no retaliation on the equilibrium path.

Inference from Reports. We note that Fact 2 also allows to identify the optimal policy from unverifiable equilibrium reports alone. Denote by $\mathbf{m}^*(\sigma)$ equilibrium reports at policy profile σ.

Fact 4. *Optimal policy σ^* solves*

$$\inf_{\sigma^N} \{\sigma_0^N \mid \mathbf{m}^*(\sigma^N) = 0 \textit{ and } \exists \sigma^O \textit{ satisfying } (4) \textit{ s.t. } \mathbf{m}(\sigma^O) = 1\}.^{15} \tag{5}$$

In words, the optimal policy is the one that requires the lowest level of baseline intervention σ_0^* consistent with: (1) message $m = 0$ being sent at σ^*; (2) message $m = 1$ being sent at an intervention profile that involves less intervention and is more informative to the agent, in the sense of exhibiting a higher likelihood-ratio of intervention rates $\frac{\sigma_1}{\sigma_0}$. Point (2) ensures that there is no silent corruption occurring at σ^* and that reports of no corruption can be trusted.

Of course, Fact 4 relies extensively on complete information and Assumption 1. We now explore the extent to which it can be extended in a general framework allowing for arbitrary incomplete information.

[15] More precisely σ^* is the limit of intervention profiles $(\sigma_n)_{n\in\mathbb{N}}$ attaining the infimum defined in (5).

3 General Framework

Our general framework relaxes the assumptions of Section 2 in three important ways: first, we allow for arbitrary incomplete information over the types of the agent and the monitor; second we allow for the possibility of covetous monitors, i.e. monitors who benefit from intervention against an honest agent; third we allow for the possibility of leaks which may reveal information over messages sent by the monitor following intervention. This allows us to identify general properties of our framework, which can be leveraged to make robust inferences over unobserved patterns of corruption, and suggest effective policy profiles.

Payoffs. Payoffs take the same general form as in Section 2, but we weaken Assumption 1 as follows.

Assumption 2 (general payoffs)**.** *There is common-knowledge that payoffs satisfy*

$$\begin{aligned} &\pi_A \geq 0; \\ \forall c \in \{0,1\}, \quad &v_A(c) \leq 0; \\ \forall c \in \{0,1\}, \quad &v_M(c, m = c) \geq v_M(c, m \neq c). \end{aligned}$$

We note that under Assumption 2, a positive mass of agents may get no benefits from corruption ($\pi_A = 0$), the certainty of intervention need not dissuade corruption ($\pi_A + v_A(c = 1) > v_A(c = 0)$), and monitors may be covetous ($v_M(c = 0, m = 1) > 0$). We continue to assume that conditional on intervention, monitors have weak preferences for telling the truth. Note that this doesn't preclude the possibility of covetous monitors, i.e. monitors that benefit from intervention happening against the agent. Consistently with this assumption, we consider policy profiles such that $\sigma_1 \geq \sigma_0$.

Information. We relax the complete information assumption of Section 2 and allow for arbitrary incomplete information. Monitors and agents have types $\tau = (\tau_M, \tau_A) \in T_M \times T_A = T$ such that the monitor's type τ_M determines her payoffs (π_M, v_M), while the agent's type τ_A determines his payoffs (π_A, v_A), and his belief over the type τ_M of the monitor, which we denote by $\Phi(\tau_M|\tau_A) \in \Delta(T_M)$. We assume that T_M is a bounded subset of $\mathbb{R}^n$.

Instead of only observing intervention, the agent now observes an abstract signal $z \in Z$ on which he can condition his retaliation policy. We assume that $z = \emptyset$ conditional on no intervention and follows some distribution $f(z|m, c)$ conditional on intervention (with $\emptyset$ remaining a possible outcome).[16]

The only restriction we impose on f is that for all $c \in \{0, 1\}$,

$$\text{prob}_f(z = \emptyset|m = 0, c) \geq \text{prob}_f(z = \emptyset|m = 1, c),$$

in words, message $m = 0$ is weakly more likely to lead to no consequences. Allowing for such general informational environments ensures that our analysis applies broadly, even if investigating institutions are not entirely trustworthy and may leak information back to the agent.

We denote by $\mu_T \in \Delta(T)$ the true distribution of types $\tau \in T$ in the population. Distribution μ_T may exhibit correlation between the types of the principal and the agent. This distribution is unknown to the principal. We think of this underlying population as a large population from which it is possible to sample independent principal agent pairs. The primary objective of Section 4 is to identify general properties of this environment, and characterize what inferences can be made on the basis of non-verifiable reports alone.

[16]For notational simplicity, we do not let this distribution depend on types, nor do we allow for differing priors over the distribution of z between the agent and the monitor. Allowing for heterogeneity in signal distribution and beliefs does not affect the analysis that follows.

4 Patterns of Corruption and Reporting

4.1 The Basic Trade-off

The basic trade-off between using information efficiently and keeping information channels open is the same as in Section 2 and Fact 1 extends without difficulty. Denote by $c^*(\sigma, \tau_A)$ the optimal corruption decision by an agent of type τ_A under policy σ, by $\mathbf{m}^*(\sigma, \tau)$ the optimal message by a monitor of type τ_M facing an agent of type τ_A under policy σ, and by $\lambda = \frac{\sigma_1}{\sigma_0}$ the likelihood ratio of intervention rates.

Proposition 1. *Assume that messages are exogenously informative, i.e. that the monitor is an automaton following strategy* $\mathbf{m}(c) = c$. *In this case, any optimal intervention profile* $\sigma^* \neq 0$ *must be such* $\sigma_0^* = 0$ *and* $\sigma_1^* > 0$.

If instead messages are endogenous, we have that

$$\liminf_{\lambda \to \infty} \int_{T_A} c^*(\sigma, \tau_A) d\mu_T(\tau_A) \geq prob_{\mu_T}(\pi_A > 0);$$
$$\forall \tau_A \ s.t. \ v_A(\cdot) < 0, \ \lim_{\lambda \to \infty} \int_{T_M} \mathbf{m}^*(\sigma, \tau) d\Phi(\tau_M | \tau_A) = 0.$$

As the likelihood ratio of intervention rates $\lambda = \frac{\sigma_1}{\sigma_0}$ gets arbitrarily large, all agents with strictly positive value for being corrupt choose to be corrupt, and all agents who suffer strictly from intervention shut down reporting (from either covetous or non-covetous monitors).

Note that the optimal intervention policy σ^* may be equal to zero if the equilibrium cost of intervention overwhelms the gains from limiting corruption. Indeed we no longer assume that certain intervention is sufficient to dissuade agents from being corrupt. As a result the principal may prefer not to intervene rather than incur high intervention costs for a marginal reduction in corruption.

4.2 The Geometry of Corruption and Reporting

Consider a given agent of type τ_A, we first show that without loss of generality, we can restrict attention to retaliation schemes that involve retaliation only conditional on intervention.

Lemma 1. *For any corruption decision c, it is optimal for the agent to retaliate only conditional on intervention: for any intervention policy σ, the agent's optimal retaliation policy is such that $r(\emptyset) = 0$.*

Indeed, retaliation conditional on $z = \emptyset$ can only increases the monitor's incentives to report the agent as corrupt. A retaliation profile $r : Z \to [0, +\infty)$ and a corruption decision c induce a messaging profile $\mathbf{m} : T_M \to \{0, 1\}$ such that for all $\tau_M \in T_M$,

$$\mathbf{m}(\tau_M) \in \arg \max_{\widehat{m} \in \{0,1\}} \sigma_{\widehat{m}}[v_M(c, \widehat{m}) - \mathbb{E}(r|c, \widehat{m})]. \tag{6}$$

We denote by $\mathcal{M} = \{0, 1\}^{T_M}$ the set of message profiles. For any corruption decision c, and any message profile $\mathbf{m} \in \mathcal{M}$, consider the measure of manipulation costs $K^{\tau_A}_{c,\mathbf{m}}(\sigma)$ defined by

$$\begin{aligned} K^{\tau_A}_{c,\mathbf{m}}(\sigma) =& \frac{1}{\sigma_0} \inf_{r:Z\to[0,+\infty)} \int_{Z\times T_M} \sigma_{\mathbf{m}(\tau_M)} k_A(r(z)) \mathrm{d}F(z|c, \mathbf{m}(\tau_M)) \mathrm{d}\Phi(\tau_M|\tau_A) \\ & \text{s.t. } \forall \tau_M,\ m = \mathbf{m}(\tau_M) \text{ satisfies,} \\ & \sigma_m \left[\mathbb{E}(v_M|m, c) - \mathbb{E}(r|m, c)\right] \geq \sigma_{\neg m} \left[\mathbb{E}(v_M|\neg m, c) - \mathbb{E}(r|\neg m, c)\right] \end{aligned} \tag{7}$$

By convention, this cost is infinite whenever message profile $\mathbf{m}$ is not implementable, i.e. when there is no retaliation profile r such that (6) holds. Noting that for all $m \in \{0, 1\}$, $\frac{\sigma_m}{\sigma_0} = \lambda^m$ and $\frac{\sigma_m}{\sigma_{\neg m}} = \lambda^{2m-1}$, it follows that the cost $K^{\tau_A}_{c,\mathbf{m}}(\sigma)$ of implementing message profile $\mathbf{m}$ can expressed as a function $K^{\tau_A}_{c,\mathbf{m}}(\lambda)$ of the likelihood ratio λ of intervention rates. The

agent will choose to be honest if and only if

$$\pi_A + \sigma_0 \sup_{\mathbf{m}\in\mathcal{M}} \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=1) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=1,\mathbf{m}}(\lambda) \right\}$$
$$\leq \sigma_0 \sup_{\mathbf{m}\in\mathcal{M}} \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=0) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=0,\mathbf{m}}(\lambda) \right\}. \tag{8}$$

This implies several useful properties for message manipulation and corruption decisions.

Proposition 2 (patterns of manipulation and corruption)**.**

(i) Pick an agent of type τ_A and consider old and new intervention profiles σ^O, σ^N such that $\sigma^O = \rho\sigma^N$, with $\rho > 0$. Denote by c^O, c^N and $\mathbf{m}^O, \mathbf{m}^N$ the corruption decisions and message profiles implemented by the agent in equilibrium at σ^O and σ^N. If $c^O = c^N$, then $\mathbf{m}^O = \mathbf{m}^N$.

(ii) Consider an agent of type τ_A. The set of intervention profiles σ such that the agent chooses to be corrupt is star-shaped around $(0,0)$: if $c^*(\sigma, \tau_A) = 1$, then $c^*(\rho\sigma, \tau_A) = 1$ for all $\rho \in [0,1]$.

(iii) Fix an intervention ratio $\lambda \geq 1$. Under the true distribution μ_T, the mass of corrupt agents

$$\int_{T_A} c^*(\sigma, \tau_A) d\mu_T(\tau_A)$$

is decreasing in baseline intervention rate σ_0.

In words, point (i) states that whenever intervention profiles have the same ratio of intervention rates, message profiles change if and only if the underlying corruption behavior of the agent is changing. Points (ii) and (iii) show that keeping the information content of intervention constant, agents are less likely to be corrupt as the intensity of intervention increases.

4.3 Inference from Unverifiable Reports

We now investigate the extent to which unverifiable reports can be used to make inferences on underlying levels of corruption and inform policy choices. Note that the only data driven observable available to the principal is the aggregate report

$$\int_T \mathbf{m}^*(\sigma,\tau)\mathrm{d}\mu_T(\tau).$$

We first highlight that in our rich environment, unverifiable messages at a single policy profile σ imply no restrictions on underlying levels of corruption.

Fact 5. *Take as given a policy profile σ, and a true distribution μ_T yielding aggregate report $\int_T \mathbf{m}^*(\sigma,\tau)d\mu_T(\tau)$. We have that*

$$\left\{\int_{T_A} c^*(\sigma,\tau_A)d\widehat{\mu}_T(\tau_A) \;\middle|\; \widehat{\mu}_T \;\; s.t \;\; \int_T \mathbf{m}^*(\sigma,\tau)d\widehat{\mu}_T(\tau) = \int_T \mathbf{m}^*(\sigma,\tau)d\mu_T(\tau)\right\} = [0,1].$$

This follows from the fact that we allow for both covetous monitors and agents who get no benefit from corruption. While reports at a single policy profile are uninformative, we now show that variation in reports across policy profiles can imply useful bounds on underlying levels of corruption.

Proposition 3. *Consider policies σ^O and σ^N such that $\sigma^N = \rho\sigma^O$, with $\rho > 1$. The following holds:*

$$\textit{(minimum honesty)} \quad \int_T [1 - c(\sigma^N,\tau_A)]d\mu(\tau_A) \geq \left|\int_T [\mathbf{m}^*(\sigma^N,\tau) - \mathbf{m}^*(\sigma^O,\tau)]d\mu_T(\tau)\right|;$$

$$\textit{(minimum corruption)} \quad \int_T c(\sigma^O,\tau_A)d\mu(\tau_A) \geq \left|\int_T [\mathbf{m}^*(\sigma^N,\tau) - \mathbf{m}^*(\sigma^O,\tau)]d\mu_T(\tau)\right|.$$

In words, changes in message patterns as policy profiles move along a ray provide lower and upper bounds to underlying levels of corruption.

Imagine that some set of policy experiments $\sigma \in \Sigma$ can be performed, where Σ is a set of feasible policy profiles. Proposition 3 suggests the following heuristic to specify intervention policies. Define $\underline{v}_P = \min_{c\in\{0,1\}} v_P(c)$, and denote by $\overline{C}$ the function such that for all $\sigma \in [0,1]^2$,

$$\overline{C}(\sigma) \equiv 1 - \max\left\{ \left| \int_T [\mathbf{m}^*(\sigma,\tau) - \mathbf{m}^*(\widehat{\sigma},\tau)] \mathrm{d}\mu_T(\tau) \right| \middle| \widehat{\sigma} \in \Sigma \cap \{\rho\sigma | \rho \in [0,1]\} \right\}.$$

From Proposition 3 we know that $\overline{C}$ is an upper bound to the amount of underlying corruption. Noting that for a given intervention profile σ, the principal's payoff is

$$\mathbb{E}_{\mu_T}[u_P|c^*,\mathbf{m}^*,\sigma] = \pi_P \int_{T_A} c^*(\sigma,\tau_A)\mathrm{d}\mu_T(\tau_A) + \int_T v_P(c^*(\sigma,\tau_A))\sigma_{\mathbf{m}^*(\sigma,\tau)}\mathrm{d}\mu_T(\tau),$$

we obtain the following corollary.

Corollary 1. *For any intervention profile σ, we have that*

$$\mathbb{E}_{\mu_T}[u_P|c^*,\mathbf{m}^*,\sigma] \geq \pi_P \overline{C}(\sigma) + \underline{v}_P \left[\sigma_0 + (\sigma_1 - \sigma_0) \int_T \mathbf{m}^*(\sigma,\tau) d\mu_T(\tau) \right].$$

Furthermore, if $\Sigma = [0,1]^2$, then the data-driven heuristic policy $\overline{\sigma}(\mu_T)$ defined by

$$\overline{\sigma}(\mu_T) \in \arg\max_{\sigma\in[0,1]^2} \pi_P \overline{C}(\sigma) + \underline{v}_P \left[\sigma_0 + (\sigma_1 - \sigma_0) \int_T \mathbf{m}^*(\sigma,\tau) d\mu_T(\tau) \right]$$

is a weakly undominated strategy with respect to the unknown true distribution μ_T.

The logic underlying policy $\overline{\sigma}(\mu_T)$ can be exploited in alternative ways which may be more practical. There are two basic steps: first, find an intervention profile that provides monitors with sufficient plausible deniability that they are willing to send complaints; second, scale up intervention rates in proportional ways until complaints diminish by a sufficient amount.

5 Discussion

5.1 Summary

We model the problem of a principal who relies on 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 which the agent can use to effectively dissuade the monitor from complaining. As a consequence, effective intervention strategies must garble the information content of messages. In particular, there needs to be a positive baseline rate of intervention following the message "non-corrupt", so as to provide the monitor with plausible deniability in the event of intervention.

To explore the extent to which one can make inferences about unobservable corruption on the basis on unverifiable messages alone, our framework allows for arbitrary heterogeneity across agents and monitors, as well as incomplete information. We establish general properties of reporting and corruption patterns which can be exploited to derive bounds on underlying corruption as a function of unverifiable reports alone. These bounds suggest heuristics to identify robust intervention policies which can be described as follows: first find intervention profiles that guarantee sufficient plausible deniability for monitors to complain, then increase intervention rates proportionally until complaints fall at an acceptable level.

The appendix extends our analysis of inference in several way. First, we explore the information content of reports in a partial equilibrium capturing short run responses to policy changes. Second, we consider different objectives for inference, such as evaluating the extent of silent corruption, i.e. corruption that is occuring but is not reported, as well as gauging the proportion of covetous monitors in the population.

5.2 Committing to mixed intervention policies

A strength of our analysis is that it does not presume that the principal has extensive control over the payoffs of the agent and the monitor. This accommodates environments in which the relevant principal may have to rely on existing institutional channels to carry out interventions, and lets us focus on a more reliably available component of policy design: the way messages are mapped to likelihood of intervention. The corresponding weakness of our analysis is that we assume the principal is able to commit to mixed strategies which is admittedly more demanding than committing to pure strategies.

One way to justify this assumption is to evaluate more closely the foundations for the principal's commitment power. We think of commitment power as arising from reputation-formation in a repeated game. Committing to mixed strategies is equivalent to forming a reputation under imperfect public monitoring. Fortunately, we know from Fudenberg and Levine (1992) that type-based reputation formation arguments hold in such settings provided that actions are statistically identifiable from signals. This is the case here since intervention can be observed by the agent.

Beyond reputation formation, we emphasize that commitment to mixed strategies can be achieved through hard-wired garbling of the messages provided by the monitor. Specifically, instead of recording messages directly, the principal may instead record the outcomes of two Bernoulli lotteries l_0 and l_1 such that

$$l_0 = \begin{cases} 1 & \text{with proba} \quad \sigma_0 \\ 0 & \text{with proba} \quad 1-\sigma_0 \end{cases} \quad \text{and} \quad l_1 = \begin{cases} 1 & \text{with proba} \quad \sigma_1 \\ 0 & \text{with proba} \quad 1-\sigma_1. \end{cases}$$

The monitor communicates by picking a lottery, with realized ouctome y. Conditional on y the principal intervenes according to pure strategy $i(y) = y$. This approach has the benefit of making plausible deniability manifest to participating monitors. Crucially, one can recover

intended aggregate reports from outcome data alone: for any mapping $m : T \to \{0, 1\}$,

$$\int_T \mathbf{m}(\tau) \mathrm{d}\mu_T(\tau) = \frac{\int_T y(\tau) \mathrm{d}\mu_T(\tau) - \sigma_0}{\sigma_1 - \sigma_0}.$$

Hence the analysis of Section 4 continues to apply as is. Note that this implementation of mixed strategies is closely related to the randomized response techniques introduced by Warner (1965).[17]

5.3 Hard versus soft measures of corruption

Our analysis has focused on inference from unverifiable messages alone. This is motivated by the fact that the principal need not have access to the outcomes of interventions, or only with significant procedural delay, and with limited reliability. Still, while it is both surprising and encouraging that one can obtain a handle on underlying corruption on the basis of unverifiable messages alone, one should not exclude the possibility of obtaining reliable direct measures of corruption.[18] In fact, soft and hard measures of corruption can usefully complement each other.

Indeed, even if the cost of obtaining hard measures of corruption limits their scalability, even a limited sample of direct measures can be used to calibrate the meaning of unverifiable reports obtained from agents. This would allow to more precisely exploit the information content of messages, better adjust intervention policies, as well as confirm or not the predictions of our analysis.

[17] The main difference is that typical randomized response techniques simply enjoin the monitor to garble his response, but the monitor can always guarantee his preferred message. Hence, in our fully rational framework, traditional randomized response techniques do not guarantee plausible deniability. This difference is important when messages are used for equilibrium incentive design, rather than for one shot surveys.

[18] See for instance Bertrand et al. (2007), Olken (2007).

Appendix

A Extensions

A.1 Short-run inference

Our analysis so far has emphasized inference in equilibrium. We now study inference under a partial equilibrium in which the monitor can adjust her behavior, while the retaliation policy of the agent remains fixed. This partial equilibrium may be more suited to interpret data collected in the short-run

We assume that corruption, retaliation and reporting policies (c^O, r^O, m^O) under policy σ^O are at equilibrium. Under the new policy σ^N, we consider the short-run partial equilibrium in which the agent's behavior is kept constant equal to c^O, r^O, while the monitor's reporting strategy m^N_{SR} best-replies to c^O, r^O under new policy σ^N.

We first note that in the short run, the policy experiment considered in Section 4 is uninformative. Indeed, consider a benchmark policy σ^O and an alternative policy σ^N such that

$$\sigma^N = \rho\sigma^O, \quad \text{with} \ \ \rho > 1.$$

Fact A.1 (no short run inferences)**.** *In the short-run equilibrium, message patterns are not affected by new policy σ^N:*

$$\forall \tau \in T, \quad \int_T m^O(\tau)\, d\mu_T(\tau) = \int_T m^N_{SR}(\tau)\, d\mu_T(\tau).$$

However, as we now show, other experimental variation may be used to extract useful information from short run data.

A lower bound on silent corruption. Consider policies σ^O and σ^N such that

$$\sigma_0^O < \sigma_0^N \quad \text{and} \quad \sigma_1^O = \sigma_1^N.$$

Proposition A.1. *Under the assumption that there are no covetous monitors and agents know it, we have that*

$$\int_T c^O(\tau_A)[1 - m^O(\tau)]d\mu_T(\tau) \geq \int_T [m_{SR}^N(\tau) - m^O(\tau)]d\mu_T(\tau).$$

A lower bound on covetousness. We now consider policies σ^O and σ^N such that

$$\sigma_0^O < \sigma_0^N \quad \text{and} \quad \sigma_1^O = \sigma_1^N.$$

Proposition A.2. *Assume that $f(z|c,m) = f(z|c)$. We have that*

$$\int_T \left[m_{SR}^N(\tau) - m^O(\tau)\right] d\mu_T(\tau) \leq \int_\tau \mathbf{1}_{V_M(c=0,m=1)>0} d\mu_T(\tau).$$

A.2 Retaliation and Side Payments

The paper relies significantly on the assumption that the agent uses retaliation to incentivize the monitor. This appendix has two objective: the first is to highlight that key results which no longer hold if the agent uses side-payments instead; the second is to endogenize the use of retaliation alone.

Comparative statics under side payments. A key result of the paper, Proposition 2 (*iii*), states that increasing intervention rates in a proportional way can only decrease corruption. This is no longer true when the agent uses rewards to provide incentives.

To make this point, it is sufficient to consider the complete information example presented

in Section 2, imposing that the agent can now only rely on rewards (which can be thought of as negative levels of retaliation). It is immediate in this setting that rewards will only be given if no intervention happens. Conditional on corruption, given a promised reward $b > 0$ following no intervention, the monitor will send message $m = 0$ if and only if

$$\begin{aligned} &(1-\sigma_0)b + \sigma_0 v_M(c=1, m=0) \geq (1-\sigma_1)b + \sigma_1 v_M(c=1, m=1) \\ \iff & b \geq \left[\frac{\sigma_1 v_M(c=1, m=1) - \sigma_0 v_M(c=1, m=0)}{\sigma_1 - \sigma_0}\right]^+ \equiv b_\sigma \end{aligned}$$

Let $k_A(b)$ denote the cost of providing reward b for the agent. For simplicity, we assume that $v_A(c=1) = v_A(c=0) = v_A$. The agent will choose to be corrupt if and only if

$$\pi_A + \max\{-(1-\sigma_0)k_A(b_\sigma) + \sigma_0 v_A, \sigma_1 v_A\} \geq \sigma_0 v_A.$$

Consider a configuration such that $-(1-\sigma_0)k_A(b_\sigma) + \sigma_0 v_A > \sigma_1 v_A$, and adjust π_A so that

$$\pi_A + \max\{-(1-\sigma_0)k_A(b_\sigma) + \sigma_0 v_A = \sigma_0 v_A - \epsilon.$$

Consider now a small increase $\Delta\sigma_0$ in σ_0, keeping $\frac{\sigma_1}{\sigma_0}$ (and therefore b_σ) constant. This diminishes the payoff from corruption by $[v_A + k_A(b_\sigma)]\Delta\sigma_0$, and diminishes the payoff from non-corruption by $v_A\Delta\sigma_0$. Hence for ϵ small enough, it follows that a proportional increase in intervention rates can increases corruption.

Sufficient conditions for the use of retaliation only. We now consider the general framework of Section 3, allow for retaliation to take negative values, i.e. $r \in \mathbb{R}$, and provide sufficient conditions for the agent to only use retaliation. The cost of retaliation k_A is extended over $\mathbb{R}$, and for simplicity, we assume that it is differentiable. Recall that state $z = \emptyset$ occurs with probability 1 if there is no intervention, and with probability $\text{prob}_f(z = \emptyset | c, m)$

if there is intervention. Let us define

$$\underline{p} = \min_{(c,m)\in\{0,1\}^2} \text{prob}_f(z = \emptyset | c, m).$$

The following holds.

Proposition A.3. *Whenever*

$$\underline{p} \times \inf_{r<0} k'_A(r) \geq (1 - \underline{p}) \times \sup_{r>0} k'_A(r),$$

for any intervention profile σ and any type τ_A, the agent's optimal retaliation strategy is such that for all z, $r > 0$, i.e. the agent uses no rewards.

Whenever the marginal cost of retaliation is low, and the probability of intervention having consequences is low, it is optimal for the agent to use retaliation only to discipline the monitor. Note that we do not assume that cost function k_A is continuously differentiable. In particular, there may be a kink at 0.

A.3 Example: the Case of Covetous Monitors

To illustrate the richness of messaging patterns that can arise when we allow for covetous monitors, we explicitly extend the example presented in Section 2 to the case where

$$v_M(c = 0, m = 1) \geq 0.$$

In words, the monitor values intervention, at least on non-corrupt agents. This may be the case if the monitor benefits from discrediting the agent, for instance she could hope to obtain the agent's position, alternatively she could benefit from replacing an honest agent with a corrupt one.

Reporting by the monitor. Take as given an intervention profile $\sigma = (\sigma_0, \sigma_1)$, with $\sigma_0 < \sigma_1$, and a level of retaliation r conditional on intervention.

When the agent is not corrupt ($c = 0$), the monitor sends message $m = 0$ if and only if

$$\sigma_1[v_M(c = 0, m = 1) - r] < \sigma_0[v_M(c = 0, m = 0) - r].$$

This holds if and only if

$$r \geq r_\sigma^0 \equiv \left[\frac{\sigma_1 v_M(c = 0, m = 1) - \sigma_0 v_M(c = 0, m = 0)}{\sigma_1 - \sigma_0}\right]^+.$$

Because the monitor is covetous, a non-corrupt agent may now have to threaten the monitor with retaliation r_σ^0 to induce the monitor to send message $m = 0$.

When the agent is corrupt, i.e. $c = 1$, the monitor will report corruption and send message $m = 1$ if and only if

$$\sigma_1[v_M(c = 1, m = 1) - r] \geq \sigma_0[v_M(c = 1, m = 0) - r].$$

This will hold whenever

$$r \leq r_\sigma^1 \equiv \left[\frac{\sigma_1 v_M(c = 1, m = 1) - \sigma_0 v_M(c = 1, m = 0)}{\sigma_1 - \sigma_0}\right]^+.$$

Note that since the monitor is covetous, we have $r_\sigma^1 \geq 0$. As before, r_σ^1 is decreasing in the ratio $\frac{\sigma_1}{\sigma_0}$. In turn $r_\sigma^0 > 0$ is decreasing in $\frac{\sigma_1}{\sigma_0}$ over the range of ratios $\frac{\sigma_1}{\sigma_0}$ such that $r_\sigma^0 > 0$. As before, the information content of intervention affects the level of retaliation needed to influence messaging.

Information manipulation and corruption. We now examine the agent's behavior. Consider the agent's incentives to manipulate information given a corruption decision $c \in$

$\{0,1\}$. Since retaliation r is costly to the agent, he either picks $r=0$ and does not influence the monitor, or picks $r=r_\sigma^c$ and induces message $m=0$ at the lowest possible cost. Hence, the agent will induce a message $\mathbf{m}(\sigma,c)$ such that

$$\mathbf{m}(\sigma,c) \in \arg\max_{m\in\{0,1\}} \sigma_m[v_A(c) - \mathbf{1}_{m=0}k_A(r_\sigma^c)]. \tag{9}$$

The agent will choose not to be corrupt if and only if

$$\pi_A + \max\{\sigma_1 v_A(c=1), \sigma_0[v_A(c=1) - k_A(r_\sigma^1)]\} \leq \max\{\sigma_1 v_A(c=0), \sigma_0[v_A(c=0) - k_A(r_\sigma^0)]\}. \tag{10}$$

The corresponding patterns of intervention, corruption and reporting are illustrated in Figure A.3. The following property holds.

Fact A.2. *There exist intervention profiles σ^O,σ^N satisfying* (4) *under which $m^O=0, c^O=1$, and $m^N=1$ and $c^N=0$.*

Optimal intervention. Similarly to the case of the non-covetous monitor, the principal's payoff is maximized either by setting $\sigma_0=\sigma_1=0$ and tolerating corruption, or by preventing corruption at the smallest possible cost, i.e. by setting the policy σ^* defined by

$$\sigma^* \in \arg\min_\sigma \sigma_{m_\sigma(c)} \quad \Big| \ \sigma \ \text{ satisfies (10)} \tag{11}$$

where $\mathbf{m}(c)$ is defined by (9).

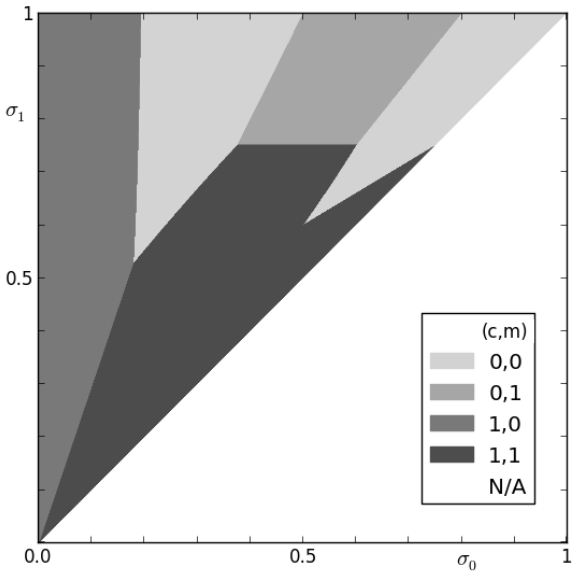


Figure 2: corruption and messages (c, m) as a function of intervention profiles (σ_0, σ_1); payoff specification $\pi_A = 3$, $v_A(c) = -6 - 4c$, $v_M(c, m) = (1 - c)(6 - m) + c(10 + 2m)$, $k_A(r) = 15r$.

B Proofs

B.1 Proofs for Section 2

Proof of Fact 1: We begin with point (i). Note that 0 is the highest payoff the principal can attain. Under intervention policy $\sigma_0 = 0$, $\sigma_1 = 1$, Assumption 1 implies that it is optimal for the agent to choose $c = 0$. As a result, there will be no intervention on the equilibrium path. Hence the principal attains her highest possible payoff, and $\sigma_0 = 0$, $\sigma_1 = 1$ is indeed an optimal intervention policy.

Let us turn to point (ii). Consider policies σ such that $\frac{\sigma_1}{\sigma_0} > 2$ and the retaliation profile under which the agent retaliates by an amount $r \equiv 2v_M(c = 1, m = 1) - v_M(c = 1, m = 0)$. Retaliation level r is chosen so that whenever the agent is corrupt, the monitor prefers to

send message $m = 0$. Indeed, the monitor prefers to send message $m = 0$ if and only if

$$\begin{aligned} &\sigma_1[v_M(c = 1, m = 1) - r] \geq \sigma_0[v_M(c = 1, m = 0) - r] \\ \iff &r \geq \frac{\lambda v_M(c = 1, m = 1) - v_M(c = 1, m = 0)}{\lambda - 1} \end{aligned} \tag{12}$$

where $\lambda = \frac{\sigma_1}{\sigma_0}$. Noting that the right-hand side of (12) is decreasing in λ and that $\lambda > 2$, we obtain that the monitor indeed sends message m whenever $r \geq v_M(c = 1, m = 1) - v_M(c = 1, m = 0)$.

It follows that a corrupt agent's expected payoff under this retaliation strategy is

$$\pi_A + \sigma_0[v_A(c = 1) - k_A(r)] \geq \pi_A + \frac{1}{\lambda}[v_A(c = 1) - k_A(r)].$$

Since $\pi_A > 0$, it follows that this strategy guarantees the agent a strictly positive payoff for λ sufficiently large. Given that the highest possible payoff for an agent choosing $c = 0$ is equal to 0, it follows that for λ large enough the agent will be corrupt.

Given corruption, we now show that the agent will also use retaliation. Under no retaliation the agent obtains an expected payoff equal to $\pi_A + \sigma_1 v_A(c = 1)$. Under the retaliation strategy described above, the agent obtains a payoff equal to $\pi_A + \frac{\sigma_1}{\lambda}[v_A(c = 1) - k_A(r)]$. Since $v_A(c = 1) < 0$ it follows that for λ large enough, it is optimal for the agent to commit to retaliation. ■

Proof of Fact 2: Let us begin with point (i). Recall that $\lambda = \frac{\sigma_1}{\sigma_0}$. We know from Section 2 that the corrupt agent induce message $m = 0$ if and only if (2) holds, i.e. if

$$\lambda v_A(c = 1) \leq v_A(c = 1) - k_A(r_\lambda).$$

From the fact that r_λ is decreasing in λ and $v_A(c = 1) < 0$, it follows that there exists λ_0

such that (2) holds if and only if $\lambda > \lambda_0$.

Consider point (ii). Note that whenever $c = 0$, since $v_A(c = 0) \leq 0$, it is optimal for the agent to never retaliate, which induces message $m = 0$. It follows that $m^N = 1$ implies $c^N = 1$. Let us define the notation $\lambda^N = \frac{\sigma_1^N}{\sigma_0^N}$ and $\lambda^O = \frac{\sigma_1^O}{\sigma_0^O}$. Since corruption is optimal for the agent at σ^N, we obtain that

$$\pi_A + \max\{\sigma_1^N v_A(c = 1), \sigma_0^N[v_A(c = 0) - k_A(r_{\lambda^N})]\} \geq 0.$$

Since $\lambda^N < \lambda^O$, r_λ is decreasing in λ, $v_A(\cdot) \leq 0$ and $\sigma^N > \sigma^O$ for the usual vector order, we obtain that

$$\pi_A + \max\{\sigma_1^O v_A(c = 1), \sigma_0^O[v_A(c = 0) - k_A(r_{\lambda^O})]\} \geq 0.$$

Hence, it must be optimal for the agent to be corrupt at σ^O: $c^O = 1$.

We now turn to point (iii). Since $m^O = 1$, we know that $c^O = 1$. Since the agent chooses not to induce message $m = 0$ at σ^O, it must be that $\lambda^O \leq \lambda_0$. Since $\lambda^N < \lambda^O$, it follows from point (i) above that a corrupt agent would not induce message $m = 0$ at σ^N. Hence, it must me that $c^N = 0$. ■

Proof of Fact 3: By Assumption 1, the optimal intervention profile must discourage corruption in equilibrium ($\sigma_0 = \sigma_1 = 1$ guarantees no corruption and is preferred to corruption in spite of high intervention costs). Since there won't be corruption in equilibrium, the equilibrium rate of intervention is σ_0. The principal's problem is therefore to find the smallest value of σ_0 for which there exists $\sigma_1 \geq \sigma_0$ satisfying

$$\pi_A + \max\{\sigma_1 v_A(c = 1), \sigma_0[v_A(c = 1) - k_A(r_\lambda)]\} \leq \sigma_0 v_A(c = 0). \tag{13}$$

Let us first show that at the optimal policy, it must be that $\sigma_1 v_A(c = 1) = \sigma_0[v_A(c =$

$1) - k_A(r_\lambda)]$. Indeed, if we had $\sigma_1 v_A(c=1) > \sigma_0[v_A(c=1) - k_A(r_\lambda)]$, then one could decrease σ_0 while still satisfying (13), which contradicts optimality. If instead we had that $\sigma_1 v_A(c=1) < \sigma_0[v_A(c=1) - k_A(r_\lambda)]$, then diminishing σ_1 increases r_λ which allows to diminish σ_0 while still satisfying (13). Hence it must be that $\sigma_1 v_A(c=1) = \sigma_0[v_A(c=1) - k_A(r_\lambda)]$. By definition of λ_0, this implies that $\sigma_1 = \lambda_0 \sigma_0$.

Hence (13) implies that $\pi_A + \sigma_1 v_A(c=1) \leq \sigma_0 v_A(c=0)$. Furthermore this last inequality must be an equality, otherwise one would again be able to diminish the value of σ_0 while satisfying (13). This implies that $\pi_A + \sigma_1 v_A(c=1) = \sigma_0 v_A(c=0)$. This proves the first part of Fact 3.

We now show that this optimal policy is necessarily interior. We know that $\sigma_0 \in (0,1)$ from Fact 1 and the assumption that $\pi_A + v_A(c=1) < v_A(c=0)$. Let us show that $\sigma_1 < 1$. The first part of Fact 3 allows us to compute σ_1 explicitly as

$$\begin{aligned}\sigma_1 &= \frac{\pi_A}{-v_A(c=1)} \frac{1}{1 - \frac{v_A(c=0)}{\lambda_0 v_A(c=1)}} \leq \frac{\pi_A}{-v_A(c=1)} \frac{1}{1 - \frac{v_A(c=0)}{v_A(c=1)}} \\ &\leq \frac{\pi_A}{-v_A(c=1) + v_A(c=0)} < 1,\end{aligned}$$

where the last inequality uses the assumption that $\pi_A + v_A(c=1) < v_A(c=0)$. This concludes the proof of Fact 3. ■

Proof of Fact 4: Fact 2 implies that any profile σ^N satisfying the condition in (5) is such that $c(\sigma^N) = 0$.

We now show that there exists a sequence of intervention profiles converging to σ^* that satisfies the conditions in (5). We know from Fact 3 that policy σ^* satisfies $\mathbf{m}^*(\sigma^*) = 0$ and

$\sigma_1^* = \lambda_0 \sigma_0^*$. Consider sequences $(\sigma_n^O)_{n\in\mathbb{N}}$ and $(\sigma_n^N)_{n\in\mathbb{N}}$ such that

$$\sigma_{0,n}^N = \left(1 + \frac{1}{n}\right)\sigma_0^* \,, \qquad \sigma_{0,n}^O = \left(1 - \frac{1}{n}\right)\sigma_0^* \,,$$
$$\sigma_{1,n}^N = \lambda_0 \left(1 - \frac{1}{n}\right)\sigma_{0,n}^N \,, \qquad \sigma_{1,n}^O = \lambda_0 \left(1 + \frac{1}{n}\right)\sigma_{0,n}^O.$$

For n sufficiently large, the pair (σ_n^O, σ_n^N) satisfies the condition in (5), and sequence $(\sigma_n^N)_{n\in\mathbb{N}}$ converges to σ^*. This concludes the proof. ■

B.2 Proofs for Section 4

Proof of Proposition 1: Consider the case where the monitor is an automaton sending exogenously informative messages $\mathbf{m}(c) = c$. We show that it is optimal to set $\sigma_0 = 0$.

Since messages are exogenous, it is optimal for the agent not to engage in retaliation regardless of his type. Therefore the agent will be corrupt if and only if

$$\pi_A + \sigma_1 v_A(c=1) \geq \sigma_0 v_A(c=0).$$

Hence we obtain that the principal's payoff is

$$\int_T \mathbf{1}_{\pi_A+\sigma_1 v_A(c=1)\geq\sigma_0 v_A(c=0)} \sigma_0 v_P(c=0) \mathrm{d}\mu_T + \int_T \mathbf{1}_{\pi_A+\sigma_1 v_A(c=1)<\sigma_0 v_A(c=0)} [\pi_P + v_P(c=1)\sigma_1] \mathrm{d}\mu_T$$
$$\leq \int_T \mathbf{1}_{\pi_A+\sigma_1 v_A(c=1)<} [\pi_P + v_P(c=1)\sigma_1] \mathrm{d}\mu_T,$$

where we used the assumption that $v_A(c) \leq 0$ for all $c \in \{0,1\}$, and $\pi_P < 0$. Hence it follows that setting σ_0 is optimal for the principal when messages are exogenously informative.

We now consider the case where messages are endogenous. A proof identical to that of Fact 1 shows that whenever $\pi_A > 0$ for λ sufficiently high, $c^*(\sigma, \tau_A) = 1$. Hence by dominated

convergence, it follows that

$$\lim_{\lambda\to\infty}\int_{T_A} c^*(\sigma,\tau_A)\int \mu_T(\tau_A) \geq \text{prob}_{\mu_T}(\pi_A > 0).$$

We now show that for all types τ_A such that $v_A(\cdot) < 0$, the agent will induce the monitor to send message $m = 0$. The proof is by contradiction. Consider an agent of type τ_A and assume that there exists $\epsilon > 0$ such that for all λ large enough,

$$\int_{T_M} \mathbf{m}^*(\sigma,\tau)\mathrm{d}\Phi(\tau_M|\tau_A) > \epsilon.$$

This implies that given a corruption decision c, the agent's payoff is bounded above by

$$\pi_A \times c + \left[\sigma_0 + (\sigma_1 - \sigma_0)\int_{T_M} \mathbf{m}^*(\sigma,\tau)\mathrm{d}\Phi(\tau_M|\tau_A)\right] v_A(c) < \pi_A \times c + \sigma_0[1 + (\lambda - 1)\epsilon]v_A(c).$$

Consider the alternative strategy in which the agent chooses corruption status c but commits to retaliate with intensity

$$r = \sup_{v_M \in \mathsf{supp}\,\Phi(\cdot|\tau_A)} [2v_M(c, m = 1) - v_M(c, m = 0)]\frac{1}{\min_{m,c}\text{prob}_f(z \neq \emptyset|m, c)}$$

whenever $z \neq 0$. This retaliation strategy ensures that all types τ_M in the support of $\Phi(\cdot|\tau_A)$ choose to send message $m = 0$. Under this strategy the agent obtains a payoff greater than

$$\pi_A \times c + \sigma_0[v_A(c) - k_A(r)].$$

For λ sufficiently large that $(\lambda - 1)v_A(c) \geq k_A(r)$, this contradicts the hypothesis that $\mathbf{m}^*$ is an optimal message manipulation strategy for the agent. Hence it must be that $\lim_{\lambda\to\infty}\int_{T_M} \mathbf{m}^*(\sigma,\tau)\mathrm{d}\Phi(\tau_M|\tau_A) = 0$. This concludes the proof of Proposition 1. ■

Proof of Lemma 1: Taking a corruption decision c as given, the agent's expected payoff under an optimal retaliation profile $\mathbf{r} : Z \to [0, +\infty)$ is

$$\begin{aligned}\pi_A \times c + \text{prob}_{\mu_T}(m = 0|\mathbf{r}, c, \sigma)\sigma_0[v_A(c) - \mathbb{E}(k_A(r)|m = 0, c)] \\ + \text{prob}_{\mu_T}(m = 1|\mathbf{r}, c, \sigma)\sigma_1[v_A(c) - \mathbb{E}(k_A(r)|m = 1, c)].\end{aligned}$$

If it is optimal for the agent to engage in a positive amount of retaliation, it must be that

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

since otherwise, no retaliation would guarantee the agent a greater payoff. We now show that setting $r(\emptyset)$ to 0 increases the probability with which the monitor sends message $m = 0$. Since it also reduces the cost of retaliation, it must increase the agent's payoff.

A monitor sends a message $m = 0$ if and only if

$$\begin{aligned}-(1 - \sigma_0)r(\emptyset) + \sigma_0[v_M(c, m = 0) - \mathbb{E}(r|m = 0, z \neq \emptyset, c)\text{prob}_f(z \neq \emptyset|m = 0, c) \\ - r(\emptyset)\text{prob}(z = \emptyset|m = 1, c)] \\ \geq -(1 - \sigma_1)r(\emptyset) + \sigma_1[v_M(c, m = 1) - \mathbb{E}(r|m = 1, z \neq \emptyset, c)\text{prob}_f(z \neq \emptyset|m = 1, c) \\ - r(\emptyset)\text{prob}(z = \emptyset|m = 1, c)].\end{aligned} \tag{14}$$

Since $\sigma_1 \geq \sigma_0$ and, by assumption, $\text{prob}_f(z \neq \emptyset|m = 1, c) \geq \text{prob}_f(z \neq \emptyset|m = 0, c)$, it follows that whenever (14) holds for a retaliation profile such that $r(\emptyset) > 0$, it continues to hold when $r(\emptyset)$ is set to 0, everything else being kept equal. Hence optimal retaliation profiles are such that $r(\emptyset) = 0$. ■

Proof of Proposition 2: We begin with point (i). We know from Section 4 that the

agent's payoff conditional on a corruption decision c and a message profile $\mathbf{m}$ can be written as

$$\pi_A \times c + \sigma_0 \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c,m}(\lambda) \right\}.$$

It follows that given a corruption decision c, the agent induces a message profile $\mathbf{m}$ that solves

$$\max_{\mathbf{m}\in\mathcal{M}} \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c,m}(\lambda).$$

Since this problem depends only on ratio $\lambda = \frac{\sigma_1}{\sigma_0}$, it follows that $\mathbf{m}^O = \mathbf{m}^N$.

Let us turn to point (ii). Assume that it is optimal for the agent to take decision $c = 0$ at intervention profile σ. It must be that

$$\begin{aligned} &\pi_A + \sigma_0 \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=1) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=1,m}(\lambda) \right\} \\ &\leq \sigma_0 \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=0) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=0,m}(\lambda) \right\}. \end{aligned}$$

Since $\pi_A \geq 0$, this implies that

$$\int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=0) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=0,m}(\lambda) - \left(\int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=1) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=1,m}(\lambda) \right) \geq 0,$$

which implies that keeping λ constant

$$\begin{aligned} &\pi_A + \sigma_0' \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=1) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=1,m}(\lambda) \right\} \\ &\leq \sigma_0' \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=0) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=0,m}(\lambda) \right\}. \end{aligned}$$

for any $\sigma_0' \geq \sigma_0$. This implies that the agent will choose not to be corrupt at any profile $\rho\sigma$, with $\rho > 1$.

Point (iii) follows from point (ii). For any σ^O, σ^N such that $\sigma^N = \rho\sigma^O$ with $\rho > 1$, we have that for all types $\tau_A \in T_A$, $c^*(\sigma^O, \tau_A) \geq c^*(\sigma^N, \tau_A)$. Integrating against μ_T yields point (iii). ■

Proof of Fact 5: Fix σ and a distribution μ_T such that $\int_T \mathbf{m}^*(\sigma, \tau)\mathrm{d}\mu_T(\tau) = M \in [0, 1]$. Fix $C \in [0, 1]$. We show that there exists $\widehat{\mu}_T$ such that $\int_T \mathbf{m}^*(\sigma, \tau)\mathrm{d}\widehat{\mu}_T(\tau) = M$ and $\int_{T_A} c^*(\sigma, \tau_A)\mathrm{d}\mu_T(\tau_A) = C$.

For simplicity we work with type spaces such that the agent knows the type of the monitor, and allow payoffs to be correlated. A possible environment is as follows. The agent observes intervention and no other signal. With probability C, the agent gets a strictly positive payoff $\pi_A > 0$ from corruption. Conditional on $\pi_A > 0$, with probability α, the monitor has high value for intervention against corrupt agents $v_M(c = 1, m) = v > 0 = v_M(c = 0, m)$; with probability $1-\alpha$, the monitor has a low value for intervention on corrupt agents: $v_M(c, m) = 0$ for all $(c, m) \in \{0, 1\}^2$. The cost of retaliation for the agent is such that k_A is convex, $k'_A(0) = 0$ and $k_A(v) = k > 0$. For $v_A(c = 1) > 0$ appropriately low, it will be optimal for the agent to be corrupt, and commit to an arbitrarily low retaliation profile so that the monitor with a low value for intervention sends message $m = 0$ and the monitor with a high value for intervention sends message $m = 1$.

With probability $1 - C$ the agent gets a payoff $\pi_A = 0$ from corruption and has an arbitrarily high cost of retaliation. The agent's values upon intervention are such that $v_A(c = 1) < v_A(c = 0)$. With probability β, the monitor has negative value for intervention against a non-corrupt agent $v_M(c = 0, m) < 0$. With probability $1 - \beta$ the monitor gets a positive payoff $v > 0$ from intervention against the agent, regardless of his corruption. For v and a cost of retaliation k_A sufficiently high, the agent will choose not to be corrupt, the non-covetous monitor will send message $m = 0$, and the covetous monitor will send message $m = 1$.

For any $C \in [0,1]$, one can find α and β such that $C\alpha + (1-C)\beta = M$. This concludes the proof. ■

Proof of Proposition 3: From Proposition 2 (ii), we obtain that $c(\sigma^O, \tau_A) - c(\sigma^N, \tau_A) \in \{0,1\}$. Using Proposition 2 (i), this implies that $c(\sigma^O, \tau_A) - c(\sigma^N, \tau_A) \geq |m(\sigma^O, \tau) - m(\sigma^N, \tau)|$. Integrating against μ_T implies that

$$\begin{aligned} &\int_T |m(\sigma^O, \tau) - m(\sigma^N, \tau)| \mathrm{d}\mu_T(\tau) \leq \int_{T_A} [c(\sigma^O, \tau_A) - c(\sigma^N, \tau_A)] \mathrm{d}\mu_T(\tau_A) \\ \Rightarrow &\left| \int_T m(\sigma^O, \tau) - m(\sigma^N, \tau) \mathrm{d}\mu_T(\tau) \right| \leq \int_{T_A} [c(\sigma^O, \tau_A) - c(\sigma^N, \tau_A)] \mathrm{d}\mu_T(\tau_A). \end{aligned}$$

Using the fact that $c(\sigma^O, \tau_A) \leq 1$ and $c(\sigma^N, \tau_A) \geq 0$, we obtain the bounds given in Proposition 3. ■

Proof of Corollary 1: The first part of the corollary follows directly from Proposition 3. The second part of the corollary follows from Fact 4. Indeed, the strategy profile $\overline{\sigma}(\mu_T)$ coincides with the optimal strategy profile whenever payoffs are complete information and Assumption 1 holds. ■

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.

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.

Brown, P. and A. Tarca (2007): "Achieving high quality, comparable financial reporting: A review of independent enforcement bodies in Australia and the United Kingdom," *Abacus*, 43, 438–473.

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. (2013): "Inside Corruption Networks: Following the Money in Community Driven Development," *Unpublished manuscript, Caltech*.

Financial Reporting Council (2004): "Policy Update," *http://www.frc.org.uk/News-and-Events/FRC-Press/Press/2004/December/Financial-Reporting-Review-Panel-Announces-2005-Ri.aspx*.

Financial Reporting Review Panel (2005–2011): "Annual Report," *http://www.frc.org.uk*.

Fudenberg, D. and D. K. Levine (1992): "Maintaining a reputation when strategies are imperfectly observed," *The Review of Economic Studies*, 59, 561–579.

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," .

Prendergast, C. (2000): "Investigating Corruption," *working paper, World Bank development group*.

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

Rahman, D. (2012): "But who will monitor the monitor?" *The American Economic Review*, 102, 2767–2797.

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," *Journal of Law, Economics, and Organizations*, 2, 181.

WARNER, S. L. (1965): "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 60, 63–69.