Corruption, Intimidation, and Whistleblowing:

A Theory of Inference from Unverifiable Reports*

Sylvain Chassang
Princeton University

Gerard Padró i Miquel†
London School of Economics

November 26, 2014.

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 an intervention. In this setting intervention policies that are very responsive to the monitor's message provide very informative signals to the agent, allowing him to shut down communication channels. Successful intervention policies must garble the information provided by monitors and cannot be fully responsive. 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, structural experiment design, prior-free policy design.

*We are grateful to Johannes Hörner for a very helpful discussion. We are indebted to Nageeb Ali, Abhijit Banerjee, Michael Callen, Yeon Koo Che, Hans Christensen, Ray Fisman, Matt Gentzkow, Bob Gibbons, Navin Kartik, David Martimort, Andrea Prat, Jesse Shapiro, as well as seminar audiences at Berkeley, Columbia, Essex, Hebrew University, the Institute for Advanced Study, the 2013 Winter Meeting of the Econometric Society, MIT, MIT Sloan, the Nemmers Prize Conference, NYU, NYU IO day, Paris School of Economics, Pompeu Fabra, 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

Organizations and regulatory agencies often attempt to protect informants and whistleblowers to improve information transmission.[1] Anonymity guarantees are widely regarded as one of the primary means to achieve this goal: the 2002 Sarbanes-Oxley act, for instance, requires public companies to establish anonymous reporting channels. However, work by Kaplan et al. (2007, 2009) shows that greater anonymity guarantees seem to have little effect on information flows in practice.[2] We believe this is explained by the fact that in many cases, the set of people informed about misbehavior is small, and formal anonymity offers little actual protection. Police officers on patrol are a particularly salient example in which anonymity becomes meaningless: misbehavior by one officer is only observed by the other.[3] In such cases, whistleblowing is easily deterred with explicit or implicit threats of retaliation from misbehaving individuals. The primary objective of this paper is to develop whistleblowing and counter-corruption policies that are effective even when the set of potential whistleblowers is small.

We formalize the problem using a principal-agent-monitor framework in which the principal relies on messages from a single informed monitor to target intervention against a potentially corrupt agent.[4] The difficulty is that the agent can credibly threaten to retaliate against the whistleblower as a function of available observables — including the principal's own intervention behavior. We show that when the monitor is exogenously truthful, policies in which the principal's decision to intervene is more responsive to the monitor's messages naturally provide greater incentives for the agent to behave well. However, when messages are endogenous, making intervention responsive to the monitor's message facilitates effec-

[1] For a review on the literature on whistleblowing from different social sciences, see Near and Miceli (1995).

[2] In Kaplan and Schultz (2007), anonymous reporting channels fail to increase intention-to-report rates relative to non-anonymous ones. Similarly, in Kaplan et al. (2009) external hotlines with stronger safeguards do not elicit a higher propensity to report than internal hotlines with weaker safeguards.

[3] Other examples include judges and courtroom officials, fraudulent firms and their external accountants, bullying or harassment in small teams, and so on ...

[4] Throughout the paper we refer to the principal and monitor as she, and to the agent as he. We refer to the agent's decisions as "corruption" and "no corruption", but these actions correspond to any decisions that the principal may find harmful or desirable.

tive retaliation by misbehaving agents and limits information provision. This generates a novel trade-off between eliciting information and using that information efficiently. In addition, this makes evaluating intervention policies difficult: imagine that no misbehavior is reported; does this imply 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 the principal's intervention rule, the agent's misbehavior, and the monitor's whistleblowing, and suggest ways to identify effective intervention strategies using only unverifiable reports.

Our modeling approach emphasizes three issues that are important in practical applications. First, we take seriously the idea that in the long run, corrupt agents will undermine the effectiveness of institutions by side-contracting with others who have information about them (i.e. the monitor). In our model, this side-contracting takes the form of contingent retaliation profiles. Second, departing from much of the literature on collusion, we do not assume that messages are verifiable, reflecting the fact that hard measures of corruption are rarely available.[5] When it comes to anti-corruption policy, the outcome of interest need not be directly measurable, and policies may have to be evaluated using non-verifiable messages. Third, we do not assume that the principal has precise control over the payoffs of either the agent or the monitor following intervention: rewards and punishment may be determined by imperfect and stochastic institutional processes; whistleblower protection schemes may not fully shield the monitor against ostracism, or harassment; supposedly anonymous information may be leaked; the judiciary may fail to act against corrupt agents, and so on.[6] In fact, we do not assume that the principal has much information about the payoffs of the monitor or the agent. We allow for rich heterogeneity in the players' payoffs, including the possibility of "malicious" monitors who benefit from triggering intervention against honest agents. As a result, the principal should be concerned that measures taken to protect whistleblowers may end up empowering scoundrels.

[5]See Bertrand et al. (2007) or Olken (2007) for innovative approaches to measuring corruption.

[6]For instance, Miceli et al. (1999) show that reported rates of retaliation against federal employees increased between 1980 and 1992 despite the tightening of whistleblower protection laws.

We provide two main sets of results. The first is that any effective intervention strategy must limit its responsiveness to the monitor's messages. Indeed, consider a principal committed to launch an intervention with high probability when receiving message "corrupt", and with low probability when the message is "not corrupt". In this case, the information content of intervention — measured by the likelihood ratio of intervention rates under the two messages — is high. Intervention itself is a signal that simplifies the agent's incentive problem vis à vis the monitor: committing to very painful retaliation conditional on intervention ensures that the monitor never reports corruption at little equilibrium cost. It follows that if the principal wants to receive informative messages from the monitor, the information content of intervention cannot be too high. In particular, the likelihood of intervention against agents reported as not corrupt must be bounded away from zero. This guarantees the monitor a form of plausible deniability that's effective in limiting retaliation even without anonymity. In addition, it may be optimal to intervene with probability less than one against agents reported as corrupt. This allows the principal to reduce intervention on non-corrupt agents, which is costly in equilibrium, and still keep the information content of intervention low. The basic take-away is that since collusion is a side-contracting problem, it can be addressed by exploiting imperfect-monitoring contracting frictions between colluding parties.

Our second set of results shows how to use equilibrium properties of corruption and reporting decisions to infer bounds on the underlying levels of corruption using non-verifiable reports alone. We show that for any given type of agent, the region of the intervention-strategy space in which corruption occurs is star-shaped around the origin. Moreover, keeping corruption behavior constant, the messages sent by monitors depend only on the information content of intervention (the ratio of intervention rates), and not on the intensity of intervention (the absolute values of intervention rates). Using these properties, we show that policy experiments which vary the level of intensity while keeping the information content of intervention constant yield useful bounds on unobservable corruption. These bounds can be used for prior-free policy design and suggest the following rule-of-thumb: first pro-

vide enough plausible deniability so that monitors are willing to complain, and then scale up intensity while keeping the information content of intervention constant.

This paper contributes 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.[7] Growing microeconomic evidence confirms the importance of corruption issues affecting public service provision and public expenditure in education or health (see Olken and Pande (2012) and Banerjee et al. (2012) for recent reviews), and recent experimental evidence suggests that appropriate incentive design can reduce misbehavior (Olken (2007), Duflo et al. (2012, 2013)). A key aspect of corruption is that although there is strong suspicion that it is occurring, there is generally little direct and actionable evidence flowing back to the relevant principals. We believe that addressing explicit or implicit threats of retaliation is essential to ensure proper information flows.[8] We show that even without anonymity, correct policy design can keep information channels open under these threats. Relying on robust implications from our structural model, we also provide a method to measure underlying corruption and guide policy elaboration using only unverifiable report data. In this respect, we contribute to a growing literature which takes a structural approach to experiment design in order to make inferences about unobservables.[9]

This paper also contributes to the contract theory literature on collusion in organizations (see for instance Tirole (1986), Laffont and Martimort (1997, 2000), Prendergast (2000), or Faure-Grimaud et al. (2003)). Our insight is that whenever collusion is an issue, then it will be in the principal's interest to make side-contracting between the agent and the monitor

[7]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.

[8]See for instance Ensminger (2013) who emphasizes the role of threats and failed information channels in recent corruption scandals affecting community-driven development projects. Also, 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 healthcare worker, and a government worker may have some power to retaliate against them."

[9]See for instance Karlan and Zinman (2009) or Chassang et al. (2012).

difficult. The forces that make contracting difficult are well known: adverse selection and moral hazard. Here we focus on the latter, and show how the principal can make the agent's own incentive provision problem more difficult by garbling the information content of the monitor's responses.[10] This creates a novel practical rationale for the use of random mechanisms, and we believe that this simple idea has applications in other settings, for instance to fight collusion in auctions.[11] Our paper also emphasizes a novel set of questions in this literature. Rather than solving for optimal contracts in a Bayesian environment, we study the inference of unobserved but payoff-relevant behavior, and the extent to which unverifiable message data can be used for prior-free policy design.[12]

Finally, our work is related to that of Myerson (1986) or more recently Rahman (2012) who consider mechanism design problems with non-verifiable reports, and emphasize the value of random recommendation-based incentives to jointly incentivize mutliple agents, and in particular to incentivize both effort provision and the costly monitoring of effort. The key difference is that this strand of literature excludes the possibility of side contracting between players. As a result, the role of mixed strategies in our work is entirely different: monitoring itself is costless and randomization occurs only to garble the information content of the principal's intervention behavior and make side-contracting between the agent and the monitor difficult.[13] Our work also 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 framework and presents the

[10]This echoes the point made by Dal Bó (2007) in a legislative context, that making votes anonymous can help prevent influence activities and vote-buying.

[11]Specifically, our approach suggests garbling the selection of the winner(s). Although there is an active theoretical and empirical literature on collusion in auctions — see, among others, Skrzypacz and Hopenhayn (2004), Che and Kim (2006, 2009) or Asker (2010) — we believe this point has yet to be made.

[12]Our frequentist data-driven approach to policy elaboration fits in growing body of work on non-Bayesian mechanism design. See for instance Hurwicz and Shapiro (1978), Segal (2003), Hartline and Roughgarden (2008), Madarász and Prat (2010), Chassang (2013), Frankel (2014), Carroll (2013).

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

main points of our analysis in the context of a simple example; Section 3 introduces our general framework; Section 4 establishes robust properties of corruption and reporting in equilibrium, and shows how they can be exploited to form estimates of underlying corruption levels as well as make policy recommendations; Section 5 concludes with a discussion of potential implementation challenges. Appendix A presents several extensions, covering the case of multiple monitors, as well as short-term out-of-equilibrium inference. We also provide suggestive anecdotal evidence for the mechanism we emphasize. Proofs are contained in Appendix B.

2 An Example

This section introduces our framework and illustrates the mechanics of corruption, intimidation and whistleblowing through a simple but detailed example. In the interest of clarity, we make restrictive assumptions which are relaxed in Sections 3 and 4.

Note that our framework makes conscious modeling choices which demand some motivation. We provide such motivation in Section 2.2 after laying out the structure of the game.

2.1 Setup

Players, timing, and actions. There are three players: a principal P, an agent A and a monitor M.[14] 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.[15]

[14] See Appendix A for an extension to the case of multiple monitors.

[15] In this simple setting, this binary message space is without loss of efficiency: collecting messages from the agent, or richer messages from the monitor (for instance about threats of retaliation) is not helpful. See Appendix B, Lemma B.1 for details.

Again, we emphasize that here "corruption" really covers any behavior that the principal finds undesirable, such as "shirking" in a typical principal-agent model.

2. The principal observes the monitor's message m and triggers an intervention ($i = 1$) or not ($i = 0$). Intervention has payoff consequences for the principal, the agent, and the monitor that are detailed below.

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

This timing of actions is associated with a specific commitment structure: the principal commits first to an intervention policy, following which the agent commits to a retaliation strategy (see further description below).

Observables and reduced-form 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 $i \in \{0, 1\}$.[16] We assume in this section (but not others) that payoffs are common-knowledge.

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

$$\begin{aligned} u_M &= \pi_M \times c + v_M(c, m) \times i - r \\ u_A &= \pi_A \times c + v_A(c) \times i - k_A(r) \\ u_P &= \pi_P \times c + v_P(c) \times i \end{aligned}$$

where π_M, π_A, and π_P capture the expected payoff consequences of corruption, v_M, v_A, and v_P capture reduced-form expected payoffs associated with intervention.[17] The level of retaliation imposed by the agent on the monitor is denoted by r, and $k_A(r)$ is the cost of such retaliation to the agent. Payoffs conditional on corruption are such that $\pi_A > 0$ and $\pi_P < 0$. The cost

[16] 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.

[17] As we discuss in Section 2.2, if the principal has some control over the rewards and punishments attributed to the agent and the monitor, these reduced-form payoffs can be thought of as endogenously arising from a first-stage optimization.

of retaliation $k_A(r)$ is strictly increasing in r, with $k_A(0) = 0$. We make the following assumption.

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

$$\begin{aligned}
\forall m \in \{0,1\}, \quad & v_M(c=0,m) < 0 && \textit{(non-malicious monitor)};\\
& \pi_A + v_A(c=1) < v_A(c=0) = 0 && \textit{(effective intervention)};\\
& \pi_P \leq v_P(c=0) < 0 && \textit{(optimality of intervention)};\\
\forall c \in \{0,1\}, \quad & v_M(c, m \neq c) \leq v_M(c, m=c) && \textit{(weak preference for the truth)};
\end{aligned}$$

The first three assumptions are made for simplicity and are relaxed in our general analysis. The assumption that there are no malicious monitors requires that the monitor gets a negative continuation payoff $v_M(c = 0, m) < 0$ following intervention on an honest agent; effective intervention requires that certain intervention does not hurt the agent if he is honest, and hurts him sufficiently when dishonest to dissuade corruption; optimality of intervention guarantees that it is always optimal for the principal to pick an intervention profile that induces the agent to be honest. The last assumption (weak preference for the truth) is maintained throughout the paper. We assume that taking intervention as given, the monitor is weakly better off telling the truth. This assumption gives an operational meaning to messages $m \in \{0, 1\}$, and typically comes for free in direct mechanism design problems.

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 either arising from reputational concerns, or being enforced by institutions. 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.[18] Without loss of generality, we focus on strategies such that $\sigma_1 \geq \sigma_0$.[19]

Knowing the principal's intervention strategy σ, the agent takes a corruption decision

[18] 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 achieved by garbling the messages provided by the monitor directly at the surveying stage, before it even reaches the principal.

[19] See Appendix B, Lemma B.1 for details.

$c \in \{0,1\}$ 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 strategic commitments of both the principal and the agent.[20] Note that we assume that retaliation, rather than payments, is the side-contracting instrument available to the agent, and this plays an important role in the analysis.[21]

We are interested in characterizing patterns of corruption and information transmission as a function of the principal's intervention policy σ. We also solve for the principal's optimal policy and show that it must be interior. For simplicity, we assume throughout the paper that whenever the agent is indifferent, he chooses to not be corrupt, and whenever the monitor is indifferent, she reveals the truth. This convention does not matter for any of our results.

2.2 Motivation

Retaliation and failure of anonymity. Our model is tailored to capture the mechanics of corruption and whistleblowing in specific settings: (1) first there must be significant information about corrupt agents which the principal wants to obtain; (2) the set individuals who have this information and are able to pass it on to the principal is small, or can be identified ex post by the agent; (3) the agent is able to retaliate (at least with some probability) even following intervention. We believe many environments exhibit these features.

Corruption may include bribe collection by state officials, arrangements between police officers or judges 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

[20] The order of moves is essential for the analysis. Intuitively, it 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: fully public commitments to retaliate would be directly incriminating.

[21] See Appendix A for a detailed discussion, as well as sufficient conditions for this to be optimal even if side-payments are available, building on the fact that rewards must be paid on the equilibrium path.

life threatened by lack of prompt support, and so on.[22] In all these cases only a few colleagues, subordinates, or frequent associates are informed about the agent's misbehavior, making anonymity ineffective. Note also that even if several monitors have information, group punishments may be used. For instance, entire communities may be denied access to public services following complaints to authorities.[23] In addition, monitors may fear that anonymity is not properly ensured and that imperfect institutions may leak the source of complaints to the agent or one of his associates. In hierarchical 360° evaluations, subordinates may not be willing to complain about their superior to their superior's boss if they worry that the two may share information.

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 instruments available to 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. This is especially true in the more general framework of Section 3.

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

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

Provided that payoffs in $\mathcal{V}$ satisfy Assumption 1 (or the more general assumption made in Section 3), our analysis applies as a second stage within the broader mechanism design

[22] See Punch (2009) for examples of punishment of informants in a study of police corruption. In the National Business Ethics Survey 2013 21% of whistleblowers report suffering several forms of retribution despite the legal and isntitutional protection available.

[23] 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.

problem in which payoffs are endogenously determined by the principal. For instance, our reduced-form payoffs capture schemes under which the monitor receives reward $v_M(c = 1, m = 1) > 0$ for correctly informing the principal that the agent is corrupt, and is instead punished for erroneous statements ($v_M(c, m \neq c) \leq 0$).

Our decision to eschew endogenizing payoffs reflects what we perceive as great heterogeneity in the ability of principals to reliably affect the payoffs of involved parties. While payoffs are a first order determinants of behavior, they are rarely available as policy instruments. Even powerful international organizations such as the World Bank need to go through local judiciary systems to target corrupt agents, which severely constrains their ability to deliver rewards and punishments. For this reason, we choose to focus on the decision to trigger intervention, in whatever form it may take, as our main policy dimension of interest.

2.3 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, there 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 the strongest 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 of intervention rates $\lambda \equiv \frac{\sigma_1}{\sigma_0}$ is high, intervention itself becomes a very informative signal of which message the monitor sent. When λ is too high, the agent can dissuade the monitor to send message $m = 1$ while keeping equilibrium retaliation 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 when she receives message $m = 0$. This gives the monitor plausible deniability when intervention takes place, and therefore makes the agent's own incentive problem vis-à-vis the monitor more costly to resolve, since retaliation must be carried out with positive probability in equilibrium.

2.4 Intervention, Reporting and Corruption

We now study in greater detail the patterns of corruption and information flow as a function of intervention policy σ. We proceed by backward induction.

Reporting by the monitor. We begin by clarifying the conditions under which the monitor will report corruption or not. Fix 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, given $c = 0$, her expected payoffs conditional on messages $m = 1$ and $m = 0$ necessarily satisfy

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

As a result, a non-corrupt agent will find it optimal to set retaliation level $r = 0$. Note that this relies on the assumption that the monitor is non-malicious ($v_M(c = 0, m = 1) \leq 0$). When the monitor is malicious ($v_M(c = 0, m = 1) > 0$), even honest agents may need to use threats 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 $x^+ \equiv \max\{x, 0\}$ by convention. Note that whenever $v_M(c = 1, m = 1) < 0$ (i.e. the monitor suffers from intervention against a corrupt agent), 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 concern in the context of foreign aid if corruption scandals cause aid to be withheld (Ensminger, 2013).

Information manipulation and corruption. We now characterize the agent's decisions. Note first that r_σ can be expressed as a function of likelihood ratio $\lambda \equiv \frac{\sigma_1}{\sigma_0}$:

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

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

Consider now the agent's incentives to influence reporting conditional on being corrupt ($c = 1$). Since retaliation r is costly to the agent, he either picks $r = 0$ and lets the monitor send her preferred message, or picks $r = r_\sigma$ and induces message $m = 0$ at the lowest possible cost. Hence, the agent will manipulate messages through the threat of retaliation if and only

if:

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

Whenever the information content of intervention λ is high enough, the agent will induce message $m = 0$, and there will be unreported corruption.

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

Altogether, the agent will choose to be corrupt if and only if

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

This can be further simplified, since by Assumption 1, $v_A(c=0) = 0$.

Optimal intervention. It is now straightforward to characterize the optimal intervention profile. One notable property is that it involves interior rates of intervention conditional on both message $m = 0$ and message $m = 1$. The reason for this is that by setting $\sigma_1 < 1$ one can lower baseline intervention rate σ_0, while keeping the likelihood-ratio of intervention $\frac{\sigma_1}{\sigma_0}$ low enough that messages from the monitor are informative.

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

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

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, and data-driven policy design. We now ask whether it is possible to make inferences about underlying corruption c on the basis of unverifiable messages m alone.

It turns out that even though messages are unverifiable and unreported corruption is a possibility, variation in messages across different policy choices provides sharp information about underlying levels of corruption.

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

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

We think of these two intervention profiles as policy experiments implemented on different subsamples of a population of agents and monitors.[24] Intervention profile σ^N involves strictly more intervention than σ^O while having a lower information content of intervention λ. As a result it may reasonably be expected to yield both less corruption and more reliable messages. Let c^O, c^N and m^O, m^N denote the respective corruption and reporting decisions in equilibrium conditional on σ^O and σ^N.

Patterns of corruption and reporting implied by conditions (2) and (3) are illustrated in Figure 1. The following result holds.

Fact 4. *For every pair of policies σ^O, σ^N satisfying* (4)

$$c^O \geq m^O; \tag{5}$$

$$c^O \geq c^N; \tag{6}$$

$$m^O > m^N \Rightarrow c^O > c^N. \tag{7}$$

In words, across policy changes that increase the frequency of intervention while also decreasing the information content of intervention: (*i*) reported corruption is always a weak underestimate of true corruption; (*ii*) the amount of underlying corruption can only diminish; and (*iii*) drops in reported corruption are a reliable indicator of drops in true corruption.

Fact 4 implies that messages and changes in messages can be used to make sharp infer-

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

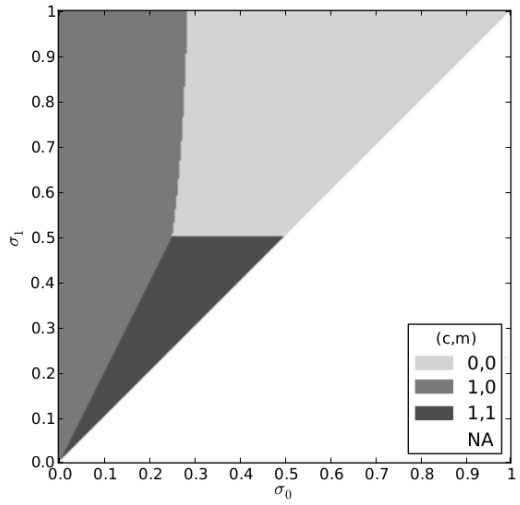


Figure 1: corruption and messages (c, m) as a function of intervention profiles (σ_0, σ_1); payoff specification $\pi_A = 5$, $v_A(c) = -10c$, $v_M(c, m) = -2 + c(3 + 3m)$, $k_A(r) = 20r$.

ences about underlying levels of corruption. A corollary is that one can identify the optimal intervention policy using unverifiable message data. Denote by $\mathbf{m}^*(\sigma)$ equilibrium reports at policy profile σ.

Fact 5. *The optimal policy σ^* solves*

$$\min_{\sigma^N}\{\sigma_0^N \mid \textit{with } \sigma^N \textit{ s.t. } \mathbf{m}^*(\sigma^N) = 0 \textit{ and } \exists \sigma^O \textit{ satisfying } (4) \textit{ s.t. } \mathbf{m}^*(\sigma^O) = 1\}. \tag{8}$$

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 frequent intervention and a higher information content of intervention λ. Point (2) ensures that there is no unreported corruption occurring at σ^* and that reports of no-corruption can be trusted.

Fragility. Some of the properties highlighted in Facts 4 and 5 are intuitive and seem like they should be robust: maintaining or increasing intervention levels, greater plausible deniability should diminish corruption; and once there is sufficient plausible deniability that

monitors report corruption, drops in reported corruption should be reliable. Unfortunately, it turns out that these useful properties do not extend to more general environments, and the following possibility results should serve as cautionary tales for policy design. Consider policies σ^O and σ^N satisfying (4). The following can happen when Assumption 1 is relaxed:

discouraging the honest – if $v_A(c = 0) < 0$, i.e. intervention is costly to an honest agent, it may be that $c^O = 0 < c^N = 1$: corruption increases with policy σ^N; this may happen if corrupt agents are being reported under σ^O, so that increasing baseline intervention rate σ_0 does not affect the payoff of corrupt agents but diminishes that of honest ones;

empowering the scoundrels – if $v_M(c = 0, m = 1) > 0$, i.e. the monitor is malicious and benefits from intervention against non-corrupt agents, it may be that $c^O < m^O$ (there is over-reporting) and that $m^O > m^N \not\Rightarrow c^O > c^N$ (drops in complaints do not imply drops in corruption); indeed, greater plausible deniability may help malicious monitors send inaccurate messages about honest agents;

unreliable drops in reports – with malicious monitors and uncertainty over payoffs, it may be that $\mathbb{E}[m^O] > \mathbb{E}[m^N]$ and $\mathbb{E}[c^O] < \mathbb{E}[c^N]$, i.e. drops in complaints are unreliable: average complaint rates can decrease while underlying levels of corruption increase.

See Appendix A for examples illustrating these different possibilities. Our general framework allows us to tackle these challenges head-on, and identify robust ways in which unverifiable messages can serve to inform policy decisions.

3 General Framework

In order to better assess what robust inferences can be drawn from our model, we generalize the framework 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 malicious 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.

Types. Payoffs take the same general form as in Section 2, but we relax the complete information assumption of Section 2 and allow for rich 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 both his payoffs (π_A, v_A, k_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 compact subset of $\mathbb{R}^n$. The only assumptions imposed on the model are the following common knowledge restriction on payoffs.

Assumption 2 (general payoffs)**.** *It 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 malicious, meaning that they benefit from intervention happening against an honest agent ($v_M(c = 0, m = 1) > 0$). 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 malicious monitors. In accordance with this assumption, we consider policy profiles such that $\sigma_1 \geq \sigma_0$.

We denote by $\mu_T \in \Delta(T)$ the true distribution of types $\tau \in T$ in the population. Distribution μ_T may exhibit arbitrary correlation between the types of the monitor and the agent, and is unknown to the principal. We think of this underlying population as a large population from which it is possible to sample independent (agent, monitor) pairs.

Leaks. We generalize the assumption that the agent can observe the principal's intervention decisions. The agent now observes an abstract signal $z \in Z \cup \{\emptyset\}$ on which he can condition his retaliation policy. We think of signal z as a potential leak from the institutional process triggered by intervention. We assume that $z = \emptyset$ conditional on no intervention and follows some distribution $F(\cdot|m, c)$ conditional on intervention. Note that $\emptyset$ remains a possible outcome conditional on intervention. In that case intervention yields no observable consequences.

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

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

that is, message $m = 0$ is weakly more likely to lead to no consequences. This ensures that in equilibrium, retaliation only occurs if intervention has been triggered. Allowing for leaks makes our analysis applicable to settings in which investigating institutions are not entirely trustworthy, resulting in information being revealed to the agent. Note that since leaks are possible, the principal has only limited commitment power and the revelation principle does not apply.[25] This is inherently an indirect mechanism design problem where messages have hard-wired institutional meaning.

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. Denote by $c^*(\sigma, \tau_A)$ the optimal corruption decision by an agent of type τ_A under policy σ, and by $\mathbf{m}^*(\sigma, \tau)$ the optimal message sent by a monitor of type τ_M facing an agent of type τ_A under policy σ. As before, let $\lambda = \frac{\sigma_1}{\sigma_0}$ denote the

[25] See Bester and Strausz (2001) for a partial extension of the revelation principle in principal-agent settings where the principal does not have commitment power. Note that in our setting leaks are not under the control of the principal.

likelihood ratio of intervention rates. Fact 1 extends as follows.

Proposition 1. *Assume that messages are exogenously informative, i.e. that the monitor is an automaton following strategy* $\mathbf{m}(c) = c$*. Any optimal intervention profile* $\sigma^* \neq 0$ *must be such that* $\sigma_0^* = 0$ *and* $\sigma_1^* > 0$.[26]

If instead messages are endogenous, we have that

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

As $\lambda = \frac{\sigma_1}{\sigma_0}$ 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 malicious or non-malicious monitors).

4.2 The Geometry of Corruption and Reporting

Consider a given agent of type τ_A. Without loss of generality, we can restrict attention to retaliation schemes that involve retaliation $r(z) > 0$ only conditional on intervention leading to some consequences, i.e. $z \neq \emptyset$.[27]

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 $\tau_M \in T_M$,

$$\mathbf{m}(\tau_M) \in \arg \max_{\widehat{m} \in \{0,1\}} \sigma_{\widehat{m}}[v_M(c, \widehat{m}) - \mathbb{E}(r | c, \widehat{m})]. \tag{9}$$

We denote by $\mathcal{M} = \{0, 1\}^{T_M}$ the set of message profiles, and for any message $m \in \{0, 1\}$ define $\neg m$ to be the other message. For any corruption decision c, and any message profile

[26] The optimal intervention policy σ^* may be equal to zero if the equilibrium cost of intervention overwhelms the gains from dissuading corruption.

[27] See Lemma B.2 for a proof.

$\mathbf{m} \in \mathcal{M}$, consider the normalized cost $K^{\tau_A}_{c,\mathbf{m}}(\sigma)$ of implementing report profile $\mathbf{m}$ defined by

$$\begin{aligned} K^{\tau_A}_{c,\mathbf{m}}(\sigma) \equiv & \frac{1}{\sigma_0} \inf_{r:Z\to[0,+\infty)} \int_{Z\times T_M} \sigma_{\mathbf{m}(\tau_M)} k_A(r(z)) \mathrm{d}F(z|c,\mathbf{m}(\tau_M)) \mathrm{d}\Phi(\tau_M|\tau_A) \\ & \text{s.t. } \forall \tau_M,\ m \equiv \mathbf{m}(\tau_M) \text{ satisfies,} \\ & \sigma_m \left[\mathbb{E}(v_M|m,c) - \mathbb{E}(r|m,c)\right] \geq \sigma_{\neg m} \left[\mathbb{E}(v_M|\neg m,c) - \mathbb{E}(r|\neg m,c)\right] \end{aligned} \tag{10}$$

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 (9) holds for all $\tau_M \in T_M$. 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. Altogether, an agent with type τ_A will choose to be honest if and only if

$$\begin{aligned} \pi_A + \sigma_0 \sup_{\mathbf{m}\in\mathcal{M}} & \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=1) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=1,\mathbf{m}}(\lambda) \right\} \\ & \leq \sigma_0 \sup_{\mathbf{m}\in\mathcal{M}} \left\{ \int_{T_M} \lambda^{\mathbf{m}(\tau_M)} v_A(c=0) \mathrm{d}\Phi(\tau_M|\tau_A) - K^{\tau_A}_{c=0,\mathbf{m}}(\lambda) \right\}. \end{aligned} \tag{11}$$

This implies several useful properties of corruption and reporting decisions in equilibrium.

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 the ratio of intervention rates $\lambda \geq 1$ and consider the ray $\{(\sigma_0, \lambda\sigma_0)$ with $\sigma_0 \in$

$[0,1]\}$. *Along this ray, 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 information content λ, message profiles change if and only if the underlying corruption behavior of the agent changes. Points (ii) and (iii) show that keeping the information content of intervention λ constant, agents are less likely to be corrupt as the intensity of intervention increases.

4.3 Inference and Policy Evaluation from Unverifiable Reports

We now investigate the extent to which unverifiable reports can be used to make inferences about the underlying levels of corruption, and to inform policy choices. Note that the only data observable to the principal is the aggregate mass of corruption messages

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

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

Proposition 3. *Take as given any interior policy profile* σ*, and a true distribution* μ_T *yielding aggregate complaint rate* $\int_T \mathbf{m}^*(\sigma, \tau) d\mu_T(\tau)$*. We have that*

$$\left\{ \int_T c^*(\sigma, \tau_A) d\widehat{\mu}_T(\tau), \quad \textit{for} \quad \widehat{\mu}_T \ \textit{s.t} \int_T \mathbf{m}^*(\sigma, \tau) d\widehat{\mu}_T(\tau) = \int_T \mathbf{m}^*(\sigma, \tau) d\mu_T(\tau) \right\} = [0,1].$$

In words, taking as given a policy profile and an observable level of aggregate complaints, one can find an underlying environment that rationalizes both the given level of complaints and any arbitrary underlying degree of corruption.

While reports at a single policy profile are uninformative, we now show that variation in

the mass of corruption messages across appropriately chosen policy profiles can imply useful bounds on the underlying levels of corruption.

Proposition 4. *Consider policies σ^O and σ^N such that $\sigma^N = \rho\sigma^O$, with $\rho > 1$. For all possible true distributions $\mu_T \in \Delta(T)$, we have that*

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

When policy profiles move along a ray, observable changes in message patterns provide a lower bound for changes in underlying levels of corruption. An immediate corollary is that changes in aggregate complaint levels $\left|\int_T [\mathbf{m}^*(\sigma^N, \tau) - \mathbf{m}^*(\sigma^O, \tau)] \mathrm{d}\mu_T(\tau)\right|$ provide a lower bound for both the mass $\int_T [1 - c^*(\sigma^N, \tau_A)] \mathrm{d}\mu_T(\tau)$ of honest agents at policy σ^N as well as a lower bound for the mass $\int_T c^*(\sigma^O, \tau_A) \mathrm{d}\mu_T(\tau)$ of corrupt agents at policy σ^O.

We now show that Proposition 4 can be used to inform policy design. Imagine that some set of policy experiments $\sigma \in \Sigma$ can be performed, where Σ is a set of feasible policy profiles. Proposition 4 suggests the following heuristic to specify intervention policies. Define $\underline{v}_P = \min_{c \in \{0,1\}} v_P(c)$, and denote by $\widehat{C} : [0,1]^2 \to [0,1]$ the function defined by

$$\forall \sigma \in [0,1]^2, \quad \widehat{C}(\sigma) \equiv 1 - \max \left\{ \left| \int_T [\mathbf{m}^*(\sigma, \tau) - \mathbf{m}^*(\sigma', \tau)] \mathrm{d}\mu_T(\tau) \right| \middle| \, \sigma' \in \Sigma \cap \{\rho\sigma | \rho \in [0,1]\} \right\}.$$

From Proposition 4 we know that $\widehat{C}(\sigma)$ is an upper bound to the amount of underlying corruption at σ. Noting that for a given intervention profile σ, the principal's payoff is

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

we obtain the following corollary.

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

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

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

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

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

While finding policy $\widehat{\sigma}(\mu_T)$ requires many policy experiments, the underlying logic can be exploited in more practical ways. The basic insight is to first, find an intervention profile with information content λ low enough that monitors are willing to send complaints; second, scale up intervention rates, keeping the information content of intervention λ constant, until complaints diminish by a sufficient amount.

5 Discussion

5.1 Summary

We model the problem of a principal who relies on messages from an informed monitor to target intervention against a possibly corrupt agent. The difficulty is that the agent can dissuade the monitor from informing the principal by threatening to retaliate conditional on observables. In this setting, intervention becomes a signal which the agent can exploit 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". This creates an imperfect monitoring problem between the agent and the monitor which limits the effectiveness with which they can side-contract.

Because hard-evidence of corruption is hard to come by, we explore the extent to which one can make inferences about unobservable corruption, as well as evaluate policies, on the basis of unverifiable messages alone. We consider a general framework wich allows for near arbitrary incomplete information and heterogeneity across agents and monitors. 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. 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 to an acceptable level.

5.2 Implementation

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 has to rely on existing institutional channels to carry out interventions. Still our policy suggestions raise some practical concerns.

Commitment to mixed strategies. Our analysis assumes 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 invoke reputational concerns in an unmodelled continuation game, committing to mixed strategies being equivalent to forming a reputation under imperfect public monitoring (Fudenberg and Levine, 1992). A more practical observation is that commitment to mixed strategies can be achieved through hard-wired garbling of messages at the surveying stage. Specifically, instead of recording messages directly, the principal may 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 privately picking a lottery, with observed realized outcome 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 aggregate submitted reports from outcome y data alone: for any mapping $\mathbf{m} : T \to \{0, 1\}$,

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

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).[28]

Destroying information. A salient concern with the policies we consider is that they require the government to explicitly garble valuable information. In particular, we show that the optimal policy may involve $\sigma_1 < 1$ and $\sigma_0 > 0$, i.e. triggering intervention against agents for whom there have been no complaints, while not investigating all agents against which a complaint has been filed. This is supect behavior that governments may prefer to avoid. This reinforces the argument that garbling should occur at the recording stage, with the caveat that the garbling procedure needs to appear natural and legitimate to participants. In addition, one may choose to focus on the subset of policies such that $\sigma_1 = 1$, so that the government cannot be suspected of abetting corruption.

Validating structural inference. Our analysis emphasizes the possibility of making structural inferences over underlying corruption on the basis of "soft" unverifiable messages. This is motivated by the fact that in many environments corruption itself is very difficult to observe. While it is encouraging that in a fairly general setting, theory allows us to place bounds on underlying corruption on the basis of unverifiable messages alone, it is legitimate to worry whether equilibrium inferences from our model can be trusted. In this respect, obtaining "hard" direct measures of corruption is valuable, even though cost limits their scalability. Indeed, even a limited sample of direct measures could be used to calibrate the meaning of unverifiable messages obtained from agents, as well as confirm or not the structural implications of our analysis.

[28]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 in all equilibria. This difference is important when messages are used for equilibrium incentive design, rather than for one-shot surveys.

Appendix - For Online Publication

A Extensions

A.1 An Anecdote

A basic implication from our paper 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. This provides the monitor with plausible deniability should her message lead to an intervention, which makes incentive provision by the agent harder.

We use anecdotal evidence from recent changes in British accounting-oversight policy to provide a plausible illustration of how this mechanism may play out in practice.[29] We emphasize that the goal here is only to describe the trade-off identified in Fact 1 and Proposition 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.[30]

Between 2004 and 2005 the UK's Financial Reporting Review Panel — the regulatory authority in charge of investigating the accounts of publicly owned firms — radically changed its investigation policy. It moved from a purely reactive policy — in which investigations were only conducted in response to complaints filed by credible agents (in our terminology, $\sigma_0 = 0$) — to a proactive policy, under which a significant number of firms were investigated each year regardless of whether complaints were filed or not (i.e. $\sigma_0 > 0$); 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.[31]

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

[30] 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.

[31] 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.

This striking pattern can be mapped to our framework as follows. 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 relationships. As a consequence, proactive investigations result in higher rates of complaints.

A.2 Examples

In this appendix we explicitly solve the model in the case where payoffs are complete information between the agent and the monitor, but allow for the more general payoffs described in Assumption 2. Specifically, intervention may be costly even to honest agents ($v_A(c = 0) < 0$) and monitors may be malicious ($v_M(c = 0, m = 1) > 0$). We describe explicitly environments for which the possibility results discussed at the end of Section 2 are true. Again, the model is solved by backward induction.

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 malicious, a non-corrupt agent may now have to threaten the monitor with positive 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

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

As before, r_σ^1 is decreasing in the ratio $\frac{\sigma_1}{\sigma_0}$. In addition r_σ^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 $m(\sigma, c)$ such that

$$m(\sigma, c) \in \arg \max_{m \in \{0,1\}} \sigma_m[v_A(c) - \mathbf{1}_{m=0} k_A(r^c_\sigma)]. \tag{12}$$

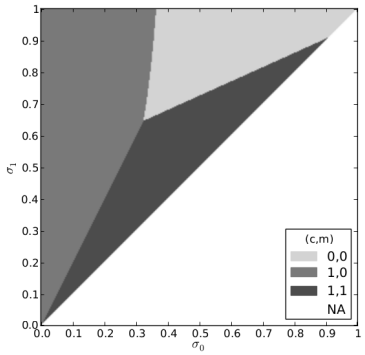
It follows that the agent will choose not to be corrupt if and only if

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

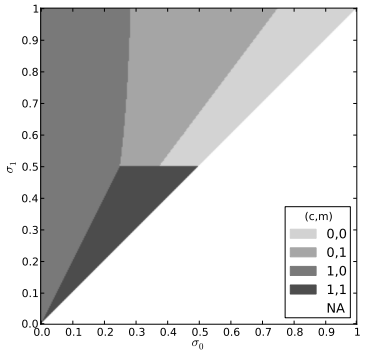
We can now provide explicit illustrations of the possibility results discussed in Section 2. Consider policies σ^O and σ^N satisfying condition (4). Figure 2(a) illustrates the fact that when $v_A(c=0) < 0$, it may be that increasing intervention (even without increasing the likelihood ratio of intervention rates) can result in greater corruption. In this example, increasing the baseline intervention rate σ_0 diminishes the payoffs of non-corrupt agents without affecting the payoffs of corrupt ones.

Figures 2(b) and 2(c) show that when the monitor is malicious, messages of corruption are no longer a lower bound to true corruption. In fact, policy changes from σ^O to σ^N can generate both increases and decreases in reports without corresponding changes in the underlying level of corruption.

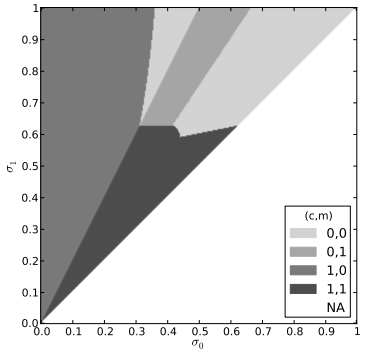
Finally, the environments of Figures 2(a) and 2(b) can be used to construct a stochastic example in which a policy change from σ^O to σ^N induces a strict drop in reported corruption and a strict increase in underlying corruption. Indeed imagine that while payoffs (v_M, v_A, π_A) are common-knowledge between the agent and the monitor, they are uncertain for the principal. In particular, say that with probability .3 payoffs are those of Figure 2(a) and with probability .7 payoffs are those of Figure 2(b). Then one can pick policies σ^O and σ^N satisfying condition (4) such that $\mathbb{E}[m^O] = .7$ and $\mathbb{E}[c^O] = 0$, while $\mathbb{E}[m^N] = .3$ and $\mathbb{E}[c^N] = .3$. In this case, a strict drop in complaints is associated with a strict increase in corruption.



(a) unproductive intervention: $v_A(c = 1) = -4.5 - 5.5c$, $v_M(c, m) = -2 + c(3 + 3m)$



(b) malicious monitor 1: $v_A(c = 1) = -10c$, $v_M(c, m) = 2 + c(-1 + 2.5m)$



(c) malicious monitor 2: $v_A(c = 1) = -2 - 8c$, $v_M(c, m) = 2 + c(-1 + 2.5m)$

Figure 2: Corruption decisions and messages (c, m) as a function of intervention profiles (σ_0, σ_1); common parameters: $\pi_A = 5$, $k_A(r) = 20r$.

A.3 Multiple Monitors

Our analysis can be extended to settings with multiple monitors. Imagine that there are now L monitors indexed by $i \in \{1, \cdots, L\}$, each of which observes the agent's corruption decision $c \in \{0, 1\}$ and can send a binary message $m_i \in \{0, 1\}$ to the principal. We denote by $\overrightarrow{m} \in \{0, 1\}^L$ the vector of messages sent by the monitors. We abuse notation and denote by 0 the message profile in which all monitors report $m_i = 0$, and by 1 the message profile in which all monitors report $m_i = 1$. An intervention policy σ is now a map $\sigma : \{0, 1\}^L \to [0, 1]$. For example, likelihood of intervention may be an affine function of the number of complaints, $\sigma_{\overrightarrow{m}} = \sigma_0 + (\sigma_1 - \sigma_0)\frac{1}{L}\sum_{i=1}^{L} m_i$. Alternatively, it may follow a threshold rule, with threshold $\Theta \in \mathbb{N}$, i.e. $\sigma_{\overrightarrow{m}} = \sigma_0 + (\sigma_1 - \sigma_0)\mathbf{1}_{\sum_{i=1}^{L} m_i > \Theta}$. For simplicity, we consider intervention policies such that for all $\overrightarrow{m}$, $\sigma_{\overrightarrow{m}} \geq \sigma_0$

As in Section 3, the agent and monitors have arbitrary types, except for the fact that Assumption 2 is common knowledge among players. We assume that each monitor i's value conditional on intervention $v_{i,M}$ depends only on c and her own message m_i. The agent now commits to a profile of vector-valued retaliation intensities $\overrightarrow{r} : Z \to [0, +\infty)^L$ associated with a cost function $k_A(\overrightarrow{r})$ that is increasing in all components of $\overrightarrow{r}$.

The vector of monitors' types is denoted by $\overrightarrow{\tau_M} = (\tau_{i,M})_{i \in \{1,\cdots,L\}}$. Note that now, each monitor's type must include a belief over other monitors' types. Furthermore, the agent's belief over $\overrightarrow{\tau_M}$ is now a joint distribution over $(T_M)^L$. Finally, the distribution of leaks z may depend on the vector of messages $\overrightarrow{m}$. We denote by $\overrightarrow{\mathbf{m}} \in \left(\{0, 1\}^{T_M}\right)^L$ profiles of message vectors as a function of the monitors' types. Note that for all $i \in \{1, \cdots, L\}$ monitor i's message profile $m_i(\tau_{i,M})$ is only a function of monitor i's type (i.e. we don't consider richer mechanisms that would let monitors' exchange information about their type).

The main properties identified in Section 4 continue to hold: for messages to be informative, it must be that all likelihood ratios of intervention rates be bounded away from 0; when policies σ are ordered along a ray, message profiles change only when corruption decisions change, and corruption must decrease along a ray going away from the origin. One difficulty is that there may now be multiple messaging equilibria among agents conditional

on a given retaliation policy. We work under the assumption that given a retaliation policy, the agent is able to select the equilibrium that most benefits him, and that this equilibrium is unique. We continue to think of the agent as selecting a message profile $\overrightarrow{\mathbf{m}}$ under constraints corresponding to the monitors' incentive compatibility conditions.

Fact A.1. *If $\sigma_0 = 0$ then all agents that benefit from corruption will be corrupt, and induce message profile $\overrightarrow{m} = 0$.*

Proof. The proof is identical to that of Proposition 1. By setting $r(z = \emptyset) = 0$ and $r(z \neq 0) = r$ arbitrarily high, the agent is able to induce message $\overrightarrow{m} = 0$ at no cost in equilibrium, which insures that there is no intervention. □

Given an interior intervention profile σ, define $\overrightarrow{\lambda} = \left(\frac{\sigma_{\overrightarrow{m}}}{\sigma_0}\right)_{\overrightarrow{m}\in\{0,1\}^L}$ the vector of likelihood ratios of intervention.

Proposition A.1. *Fix a vector of intervention ratios $\overrightarrow{\lambda}$ and consider the ray of intervention policies $\{\sigma_0 \overrightarrow{\lambda}$ for $\sigma_0 \in [0,1]\}$. Along this ray the following properties hold:*

(i) conditional on a corruption decision c, the message profile $\overrightarrow{\mathbf{m}}$ that a given agent chooses to induce is constant along the ray;

(ii) the agent's decision to be corrupt is decreasing in σ_0 along the ray.

Proof. The proof is essentially identical to that of Proposition 2. Let us begin with point (i). Conditional on a corruption decision $c \in \{0,1\}$, for any message profile $\overrightarrow{\mathbf{m}}$, we define the cost $K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}(\sigma)$ of inducing message profile $\overrightarrow{\mathbf{m}}$ as

$$\begin{aligned}
K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}(\sigma) =& \frac{1}{\sigma_0} \inf_{r:Z\to[0,+\infty)} \iint_{Z\times T_M^L} \sigma_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} k_A(r(z)) \mathrm{d}F(z|c, \overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})) \mathrm{d}\Phi(\overrightarrow{\tau_M}|\tau_A) \\
& \text{s.t. } \forall \overrightarrow{\tau_M} = (\tau_{i,M})_{i\in\{1,\cdots,L\}}, \ \ (m_i)_{i\in\{1,\cdots,L\}} = \overrightarrow{\mathbf{m}}(\tau_{i,M}) \text{ satisfies} \\
& \forall i \in \{1,\cdots,L\}, \\
& \mathbb{E}\left[\sigma_{(m_i,\overrightarrow{\mathbf{m}}_{-i})} v_{i,M}(m_i,c) - r_i \middle| m_i, \overrightarrow{\mathbf{m}}_{-i}, c\right] \geq \mathbb{E}\left[\sigma_{(\neg m_i,\overrightarrow{\mathbf{m}}_{-i})} v_{i,M}(\neg m_i,c) - r_i \middle| \neg m_i, \overrightarrow{\mathbf{m}}_{-i}, c\right].
\end{aligned}$$

It follows from inspection that $K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}$ is a function of $\overrightarrow{\lambda}$ only. By convention $K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}$ is set to $+\infty$ whenever message profile $\overrightarrow{\mathbf{m}}$ is not implementable. Given a corruption decision c, the agent chooses to induce the message profile $\overrightarrow{\mathbf{m}}$ solving

$$\sigma_0 \max_{\overrightarrow{\mathbf{m}}} \left\{ \mathbb{E}\left[\overrightarrow{\lambda}_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} v_A(c) \right] - K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}(\overrightarrow{\lambda}) \right\}.$$

It follows that the optimal message induced by the agent is a function of $\overrightarrow{\lambda}$ only, and therefore remains constant along rays.

We now turn to point (ii). An agent chooses to be non-corrupt if and only if

$$\pi_A + \sigma_0 \max_{\overrightarrow{\mathbf{m}}} \left\{ \mathbb{E}\left[\overrightarrow{\lambda}_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} v_A(1) \right] - K^{\tau_A}_{1,\overrightarrow{\mathbf{m}}}(\overrightarrow{\lambda}) \right\} \leq \sigma_0 \max_{\overrightarrow{\mathbf{m}}} \left\{ \mathbb{E}\left[\overrightarrow{\lambda}_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} v_A(0) \right] - K^{\tau_A}_{0,\overrightarrow{\mathbf{m}}}(\overrightarrow{\lambda}) \right\}. \quad (14)$$

Since $\pi_A \geq 0$ it follows that whenever (14) holds for σ_0, it must also hold for all $\sigma_0' \geq \sigma_0$. This proves point (ii). □

An implication is that changes in reporting patterns along a ray can be assigned to changes in corruption. Consider two policies σ^O, σ^N such that $\overrightarrow{\lambda}^O = \overrightarrow{\lambda}^N = \overrightarrow{\lambda}$ and $\sigma_0^O < \sigma_0^N$. For any function $X : \overrightarrow{m} \in \{0,1\}^L \mapsto x \in \mathbb{R}^n$ computing a summary statistic of messages, denote by $\widehat{\mu}^\sigma_X$ the distribution over $x \in X(\{0,1\}^L)$ defined by $\widehat{\mu}^\sigma_X(x) = \int_T \mathbf{1}_{X(\overrightarrow{m}^*(\sigma,\tau))=x} \mathrm{d}\mu_T(\tau)$, where $\overrightarrow{m}^*(\sigma,\tau)$ is the equilibrium vector of messages for a realized profile of types $\tau = (\tau_A, \overrightarrow{\tau}_M)$ given intervention policy σ. Given policies σ^O, σ^N, let D denote the distance between message distributions induced by σ^O and σ^N defined by $D \equiv \frac{1}{2}\sum_{x \in X(\{0,1\}^L)} |\mu_X^{\sigma^N}(x) - \mu_X^{\sigma^O}(x)|$. Note that D can be computed from message data alone. Proposition 4 extends as follows.

Proposition A.2 (inference)**.** *For all possible true distributions μ_T, we have that*

$$\int_T \left[c^*(\sigma^O, \tau_A) - c^*(\sigma^N, \tau_A) \right] d\mu_T(\tau) \geq D$$

which implies that D is a lower bound for the mass $\int_T [1 - c^*(\sigma^N, \tau_A)] d\mu_T(\tau)$ of honest agents at policy σ^N as well as a lower bound for the mass $\int_T c^*(\sigma^O, \tau_A) d\mu_T(\tau)$ of corrupt agents at

policy σ^O.

Proof. The proof is essentially identical to that of Proposition 4. From Proposition A.1, it follows that

$$\begin{aligned}
\int_T [c^*(\sigma^O,\tau_A)-c^*(\sigma^N,\tau_A)]\mathrm{d}\mu_T(\tau) &\geq \int_T \mathbf{1}_{\overrightarrow{\mathbf{m}_{\tau_A}}^*(\sigma^O)\neq\overrightarrow{\mathbf{m}_{\tau_A}}^*(\sigma^N)}\mathrm{d}\mu_T(\tau_A)\\
&\geq \int_{T_A}\max_{x\in X}\left\{\mathbf{1}_{\int_{\tau_M}\mathbf{1}_{X\left(\overrightarrow{\mathbf{m}_{\tau_A}}^*(\sigma^O,\overrightarrow{\tau_M})\right)=x}\mathrm{d}\mu_T(\tau_M|\tau_A)\neq\int_{\tau_M}\mathbf{1}_{X\left(\overrightarrow{\mathbf{m}_{\tau_A}}^*(\sigma^N,\overrightarrow{\tau_M})\right)=x}\mathrm{d}\mu_T(\tau_M|\tau_A)}\right\}\mathrm{d}\mu_T(\tau_A)\\
&\geq \frac{1}{2}\int_{T_A}\sum_{x\in X}\left|\int_{\tau_M}\mathbf{1}_{X\left(\overrightarrow{\mathbf{m}_{\tau_A}}^*(\sigma^O,\overrightarrow{\tau_M})\right)=x}\mathrm{d}\mu_T(\tau_M|\tau_A)-\int_{\tau_M}\mathbf{1}_{X\left(\overrightarrow{\mathbf{m}_{\tau_A}}^*(\sigma^N,\overrightarrow{\tau_M})\right)=x}\mathrm{d}\mu_T(\tau_M|\tau_A)\right|\mathrm{d}\mu_T(\tau_A)\\
&\geq D
\end{aligned}$$

which concludes the proof. □

A.4 Retaliation and Side Payments

The paper assumes that the agent uses only retaliation to provide incentives to the monitor. It is immediate that the analysis of Section 4 can be extended to allow for side-payments (modeled as $r(z) < 0$), provided that there are no rewards given conditional on no-intervention, i.e. provided that $r(z = \emptyset) = 0$.

We now provide sufficient conditions for this to be true in the general framework of Section 3, even if we allow for retaliation as well as side payments, i.e. $r \in \mathbb{R}$. The cost of retaliation $k_A(\cdot) \geq 0$ is extended over $\mathbb{R}$. For simplicity, we assume that k_A is everywhere differentiable, except at $r = 0$, where there is a kink: $k'_A(0^-) < 0 \leq k'_A(0^+)$. Recall that state $z = \emptyset$ occurs with probability 1 if there is no intervention, and with probability $\mathrm{prob}_F(z = \emptyset|c, m)$ if there is intervention. Let us define $\underline{p} = \min_{(c,m)\in\{0,1\}^2}\mathrm{prob}_F(z = \emptyset|c, m)$. The following holds.

Proposition A.3. *Whenever*

$$-\underline{p}\times\sup_{r<0}k'_A(r) > (1-\underline{p})\times\sup_{r>0}k'_A(r) > 0, \tag{15}$$

for any intervention profile σ *and any type* τ_A*, the agent's optimal retaliation strategy is such that* $r(\emptyset) = 0$.

Whenever the marginal cost of retaliation is low and the probability of intervention yielding additional information is low, it is optimal for the agent never to give out rewards when there are no observed consequences, i.e. $z = \emptyset$. Note that it may still be optimal for the agent to give out rewards, for instance if he gets a particularly informative signal z that the monitor sent message $m = 0$.

Proof. By an argument identical to that of Lemma B.2 (see Appendix B), it follows that at any optimal retaliation profile, $r(\emptyset) \leq 0$. Assume that $r(\emptyset) < 0$. We show that for $\epsilon > 0$ small enough, it is welfare improving for the agent to reduce rewards by ϵ conditional on $z = \emptyset$, and increase retaliation by ϵ at all states $z \neq \emptyset$, i.e. to use retaliation policy $r^\epsilon(\cdot) \equiv r(\cdot) + \epsilon$.

We first show that this change in retaliation policy induces the same messages from monitors. This is immediate since payoffs have been shifted by a constant: for all $m \in \{0, 1\}$, we have

$$-(1-\sigma_m)r^\epsilon(\emptyset) + \sigma_m \left[v_M(c,m) - \mathbb{E}(r^\epsilon | i = 1, m, c)\right] = -(1-\sigma_m)r(\emptyset) + \sigma_m \left[v_M(c,m) - \mathbb{E}(r | i = 1, m, c)\right] - \epsilon,$$

which implies that the monitor's IC constraints are unchanged, and retaliation profile r^ϵ induces the same message profile as r.

We now show that using r^ϵ rather than r reduces the agent's expected retaliation costs. Indeed, the change in the agent's retaliation costs is given by

$$\iint_{T_M \times Z} [k_A(r^\epsilon(z)) - k_A(r(z))] f(z | m^*(\tau_M), c) \mathrm{d}z \mathrm{d}\Phi_A(\tau_M) \leq \epsilon \left[\underline{p} \sup_{r<0} k_A'(r) + (1-\underline{p}) \sup_{r>0} k_A'(r)\right] < 0$$

where we used condition (15). This implies that it is not optimal for the agent to choose a retaliation strategy such that $r(\emptyset) < 0$. □

A.5 Short-run inference

The analysis of Section 4 emphasized inference in equilibrium. We now study inference under a partial equilibrium in which the monitor adjusts her behavior, while the corruption and

retaliation decisions of the agent remain fixed. It is plausible that this partial equilibrium model may be better 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} is a best-reply to c^O, r^O under the new policy σ^N.

We first note that in the short run, the policy experiment considered in Section 4 is uninformative.

Fact A.2 (no short-run inferences)**.** *Consider policies σ^O and σ^N such that $\sigma^N = \rho\sigma^O$ with $\rho > 1$. In the short-run equilibrium, message patterns are not affected by new policy σ^N:*

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

Proof. The result follows from the fact that given a retaliation strategy, the monitor's reporting decision, described by (9), depends only on the likelihood ratio of intervention rates. □

This is not necessarily a negative result: it can serve as a test of whether the short-run or long-run equilibrium is most suited for analysis. It also implies that the bounds given in Proposition 4 remain valid if players play a mixture of long-run and short-run equilibria. We now show that under additional assumptions, other experimental variation may be used to extract useful information from short-run data. We first describe variation useful to place bounds on unreported corruption.

Proposition A.4 (a lower bound on unreported corruption)**.** *Consider policies σ^O and σ^N such that $\sigma^O_0 < \sigma^N_0$ and $\sigma^O_1 = \sigma^N_1$. Under the assumption that there are no malicious monitors and agents know it, we have that*

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

In words, the increase in reports is a lower bound for the amount of unreported corruption.

Proof. The fact that there are no malicious monitors and the agent knows it implies that conditional on being non-corrupt, i.e. choosing $c = 0$, the agent never threatens to retaliate, i.e. $r(\cdot) = 0$. In addition, since there are no malicious monitors, it must be that $m^N_{SR}(\tau) = 1$ implies $c^O(\tau_A) = 1$. As a consequence, whenever $m^N_{SR}(\tau) - m^O(\tau) > 0$, it must be that $m^O(\tau) = 0$ and $c^O(\tau_A) = 1$. Therefore $c^O(\tau_A)(1 - m^O(\tau)) \geq m^N_{SR}(\tau) - m^O(\tau)$. This concludes the proof. □

We now describe variation allowing to obtain a bound on the number of malicious monitors.

Proposition A.5 (a lower bound on the mass of malicious monitors)**.** *Assume that there are no leaks, i.e.* $f(z|c = 1, m) = f(z|c = 1)$. *Consider policies* σ^O *and* σ^N *such that* $\sigma^O_0 = \sigma^N_0$ *and* $\sigma^O_1 < \sigma^N_1$. *We have that*

$$\int_T \mathbf{1}_{v_M(c=0,m=1)>0} d\mu_T(\tau) \geq \int_T \left[m^N_{SR}(\tau) - m^O(\tau)\right] d\mu_T(\tau).$$

Proof. Whenever $m^N_{SR}(\tau) - m^O(\tau) > 0$, it must be that $m^N_{SR}(\tau) = 1$ and $m^O(\tau) = 0$. We show that this can only occur when $c = 0$ and $v_M(c = 0, m = 1) > 0$.

Indeed, consider first the case where $c = 0$ and $v_M(c = 0, m = 1) \leq 0$. The fact that $m^O(\tau) = 0$ implies that

$$\sigma^O_1 \left[v_M(c = 0, m = 1) - \mathbb{E}(r|m = 1, c = 0)\right] \leq \sigma^O_0 \left(v_M(c = 0, m = 0) - \mathbb{E}[r|m = 0, c = 0]\right).$$

Since in this case $v_M(c = 0, m = 1) - \mathbb{E}[r|m = 1, c = 0] \leq 0$, the fact that $\sigma^N_1 > \sigma^O_1$ and $\sigma^N_0 = \sigma^O_0$ implies

$$\sigma^N_1 \left[v_M(c = 0, m = 1) - \mathbb{E}(r|m = 1, c = 0)\right] \leq \sigma^N_0 \left[v_M(c = 0, m = 0) - \mathbb{E}(r|m = 0, c = 0)\right].$$

Hence we also obtain that $m^N_{SR}(\tau) = 0$.

Consider now the case where $c = 1$. Using the fact that $f(z|c = 1, m = 1) = f(z|c =$

$1, m = 0)$, the fact that $m^O(\tau) = 0$ implies that

$$\sigma_1^O \left[v_M(c = 1, m = 1) - \mathbb{E}(r|c = 1])\right] \leq \sigma_0^O \left(v_M(c = 1, m = 0) - \mathbb{E}[r|c = 1]\right). \tag{16}$$

By Assumption 2, we have that $v_M(c = 1, m = 1) \geq v_M(c = 1, m = 0)$. Given that $\sigma_1^O > \sigma_0^O$, condition (16) can only hold if $v_M(c = 1, m = 1) - \mathbb{E}[r|c = 1] \leq 0$. This implies that necessarily,

$$\sigma_1^N \left[v_M(c = 1, m = 1) - \mathbb{E}(r|c = 1])\right] \leq \sigma_0^N \left(v_M(c = 1, m = 0) - \mathbb{E}[r|c = 1]\right).$$

We have now established that $m_{SR}^N(\tau) - m^O(\tau) > 0$ implies $v_M(c = 0, m = 1) > 0$. Proposition A.5 follows by integration over $\tau \in T$. □

B Proofs

B.1 Proofs for Section 2

We begin by establishing the simplifying claims made in Section 2.

Lemma B.1. *It is without loss of efficiency for the principal to: (i) not elicit messages from the agent; (ii) offer the monitor only binary messages* $0, 1$*; (iii) use an intervention policy satisfying* $\sigma_0 \leq \sigma_1$*.*

Proof. We begin by showing point (i): it is without loss of efficiency not to elicit messages from the agent. The agent has commitment power and therefore can commit to the messages he sends. When the agent sends a message, we can think of him as choosing the intervention profile σ that he will be facing, as well as the messages sent by the monitor. If a non-corrupt agent chooses intervention profile σ, then giving additional choices can only increase the payoffs of a corrupt agent. Hence the principal can implement the same outcome by offering only the profile σ chosen by a non-corrupt agent.

We now turn to point (ii) and consider enlarging the set of messages submitted by the monitor. The monitor observes only two pieces of information: the corruption status $c \in \{0,1\}$ of the agent, and the level of retaliation $r \in \mathbb{R}$ that he is threatened with in the event of intervention. A priori, the principal may elicit messages $(m,\rho) \in \{0,1\} \times [0,+\infty)$ about both the corruption status of the agent and the retaliation level she has been threatened with. This means that intervention rates now take the form $\sigma_{m,\rho} \in [0,1]$.

Take as given an intervention profile $\sigma = (\sigma_{m,\rho})_{m\in\{0,1\},\rho\in[0,+\infty)}$. First, note that we can focus on the case where the agent's optimal decision is to be non-corrupt, otherwise no-intervention is the optimal policy. Second, noting that the value of ρ submitted by the monitor must solve $\max_{\rho\in[0,+\infty)} \sigma_{m,\rho}(v_M(c,m) - r)$ it follows that without loss of generality one can focus on binary values of $\rho \in \{-,+\}$ such that $\sigma_{m,-} = \inf_{\rho\in[0,+\infty)} \sigma_{m,\rho}$ and $\sigma_{m,+} = \sup_{\rho\in[0,+\infty)} \sigma_{m,\rho}$.[32] Finally, without loss of efficiency, one can consider intervention profiles such that for all $\rho \in \{-,+\}$, $\sigma_{0,\rho} \leq \sigma_{1,\rho}$. Indeed, given ρ, define $\overline{\sigma} = \max_{m\in\{0,1\}} \sigma_{m,\rho}$ and $\underline{\sigma} = \min_{m\in\{0,1\}} \sigma_{m,\rho}$, as well as $\underline{m}$ and $\overline{m}$ the corresponding messages. Given ρ, the level of retaliation needed to induce $\underline{\sigma}$ rather than $\overline{\sigma}$ must satisfy

$$\overline{\sigma}(v_M(c,\overline{m}) - r) \leq \underline{\sigma}(v_M(c,\underline{m}) - r) \iff r \geq \left[\frac{\overline{\sigma}v_M(c,\overline{m}) - \underline{\sigma}v_M(c,\underline{m})}{\overline{\sigma} - \underline{\sigma}}\right]^+.$$

setting $\overline{m} = 1$ and $\underline{m} = 0$ maximizes the cost of inducing $\underline{\sigma}$ for the corrupt agent and minimizes the cost of inducing $\underline{\sigma}$ for the non corrupt agent. Note that this proves point (iii).

Given a profile σ satisfying the properties established above, we now establish the existence of a binary intervention profile $\widehat{\sigma} = (\widehat{\sigma}_m)_{m\in\{0,1\}}$ which keeps the payoff of a non-corrupt agent the same and can only decrease the payoff of a corrupt agent. Specifically set $\widehat{\sigma}_0 = \sigma_{0,-}$ and set $\widehat{\sigma}_1$ as the intervention rate that would occur under σ if a corrupt agent chose retaliation level $r = 0$. First note that the assumption that the monitor is not malicious implies that a non-corrupt agent will induce intervention rate $\sigma_{0,-}$ without using retaliation under both σ and $\widehat{\sigma}$. Hence the payoff of a non-corrupt agent remains unchanged, and the equi-

[32] When the monitor is indifferent, she must be inducing the lowest possible intervention rate, otherwise the agent would increase retaliation by an arbitrarily small amount.

librium intervention rate remains the same in both settings. Consider now the problem of the corrupt agent under $\widehat{\sigma}$. The respective costs of inducing intervention rates $\sigma_{0,-}$ and $\widehat{\sigma}_1$ haven't changed. However the agent now has less choice regarding the intervention rates she can induce. It follows that the corrupt agent must be weakly worse-off. Hence profile $\widehat{\sigma}$ also induces the agent to be non-corrupt. This concludes the proof.

□

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

$$\begin{aligned} &\sigma_1[v_M(c = 1, m = 1) - r] \geq \sigma_0[v_M(c = 1, m = 0) - r] \\ \iff &r \geq \frac{\lambda v_M(c = 1, m = 1) - v_M(c = 1, m = 0)}{\lambda - 1} \end{aligned} \tag{17}$$

where $\lambda = \frac{\sigma_1}{\sigma_0}$. Noting that the right-hand side of (17) is decreasing in λ and that $\lambda > 2$, we obtain that the monitor indeed sends message m whenever $r \geq 2v_M(c = 1, m = 1) - v_M(c = 1, m = 0)$.

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: Recall that $\lambda = \frac{\sigma_1}{\sigma_0}$. As shown in the text, the corrupt agent induces 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 $\lambda > \lambda_0$. ■

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

$$\pi_A + \max\{\sigma_1 v_A(c=1), \sigma_0[v_A(c=1) - k_A(r_\lambda)]\} \leq \sigma_0 v_A(c=0). \tag{18}$$

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 (18), 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 (18). 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 (18) 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 (18). 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: Condition (5) follows from the proof the fact that since there are no malicious monitors, an non-corrupt agent can induce message $m=0$ at no retaliation cost.

Let us now show that if $c^N = 1$ then $c^O = 1$. It follows from (5) that $m^N = 1$ implies $c^N = 1$. Let us define $\lambda^N \equiv \frac{\sigma_1^N}{\sigma_0^N}$ and $\lambda^O \equiv \frac{\sigma_1^O}{\sigma_0^O}$. Assume that $c^N = 1$. 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: $c^O = 1$.

Finally, we prove (7). 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 $\lambda^O \leq \lambda_0$, where λ_0 was defined in Fact

2. 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 be that $c^N = 0$. ■

Proof of Fact 5: Fact 4 implies that any profile σ^N satisfying the condition in (8) is such that $c(\sigma^N) = 0$.

We now show that there exists a sequence of intervention profiles converging to optimal policy σ^* that satisfies the conditions in (8). 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 (8), 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

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

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 $\sigma_0 = 0$ 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

$$\lim_{\lambda\to\infty}\int_T c^*(\sigma,\tau_A)\mathrm{d}\mu_T(\tau) \geq \mathrm{prob}_{\mu_T}(\pi_A > 0).$$

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 λ arbitrarily large,

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

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

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

whenever $z \neq \emptyset$. 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 $(\lambda - 1)\epsilon v_A(c) \geq k_A(r)$, 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. ■

Lemma B.2. *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 $r(\emptyset) = 0$.*

Proof of Lemma B.2: Taking a corruption decision c as given, the agent's expected payoff under an optimal retaliation profile $r : Z \to [0, +\infty)$ is

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

Therefore, 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

$$\begin{aligned}
&-(1-\sigma_0)r(\emptyset) + \sigma_0[v_M(c, m=0) - \mathbb{E}(r|m=0, z\neq\emptyset, c)\text{prob}_F(z\neq\emptyset|m=0,c) \\
&\qquad - r(\emptyset)\text{prob}(z=\emptyset|m=1,c)] \\
&\qquad \geq -(1-\sigma_1)r(\emptyset) + \sigma_1[v_M(c, m=1) - \mathbb{E}(r|m=1, z\neq\emptyset, c)\text{prob}_F(z\neq\emptyset|m=1,c) \\
&\qquad\quad - r(\emptyset)\text{prob}(z=\emptyset|m=1,c)].
\end{aligned} \tag{19}$$

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 (19) 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

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

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 Proposition 3: 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 $\int_T c^*(\sigma, \tau_A) \mathrm{d}\mu_T(\tau) = C$.

It is sufficient to work with type spaces such that the agent knows the type of the monitor, provided we 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 positive value for intervention against corrupt agents, i.e. $v_M(c=1,m) = v > 0 = v_M(c=0,m)$; 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 and strictly increasing. 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 complementary 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 status. For v and a cost of retaliation k_A sufficiently high, the agent will choose not to be corrupt, the non-malicious monitor will send message $m = 0$, and the malicious monitor will send message $m = 1$.

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

Proof of Proposition 4: 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

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

This concludes the proof. ■

Proof of Corollary 1: The first part of the corollary follows directly from Proposition 4. The second part of the corollary follows from Fact 5. Indeed, the strategy profile $\widehat{\sigma}(\mu_T)$ 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.

Asker, J. (2010): "A study of the internal organization of a bidding cartel," *The American Economic Review*, 724–762.

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

Banerjee, A., S. Mullainathan, and R. Hanna (2012): "Corruption," in *Handbook of Organizational Economics*, ed. by R. Gibbons and J. Roberts, Princeton University Press.

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.

Bester, H. and R. Strausz (2001): "Contracting with imperfect commitment and the revelation principle: the single agent case," *Econometrica*, 69, 1077–1098.

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.

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

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

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

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

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

Dal Bó, E. (2007): "Bribing voters," *American Journal of Political Science*, 51, 789–803.

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

Duflo, E., R. Hanna, and S. P. Ryan, N. (2012): "Incentives work: Getting teachers to come to school," *The American Economic Review*, 102, 1241–1278.

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*.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort (2003): "Collusion, delegation and supervision with soft information," *The Review of Economic Studies*, 70, 253–279.

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*.

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

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," .

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

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

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

Kaplan, S. E., K. Pany, J. A. Samuels, and J. Zhang (2009): "An Examination of the Effects of Procedural Safeguards on Intentions to Anonymously Report Fraud," *Auditing: A Journal of Practice & Theory*, 28, 273–288.

Kaplan, S. E. and J. J. Schultz (2007): "Intentions to Report Questionable Acts: An Examination of the Influence of Anonymous Reporting Channel, Internal Audit Quality, and Setting," *Journal of Business Ethics*, 71, 109–124.

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

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

Laffont, J.-J. and D. Martimort (2000): "Mechanism design with collusion and correlation," *Econometrica*, 68, 309–342.

Madarász, K. and A. Prat (2010): "Screening with an Approximate Type Space," *Working Paper, London School of Economics*.

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

Miceli, M. P., M. Rehg, J. P. Near, and C. C. Ryan (1999): "Can Laws Protect Whistle-Blowers? Results of a Naturally Occuring Field Experiment," *Work and Occupations*, 26, 129–151.

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

Near, J. and M. P. Miceli (1995): "Effective Whistleblowing," *Academy of Management Review*, 679–708.

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 (2012): "Coruption in Developing Countries," *Annual Review of Economics*, 4, 479–505.

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.

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

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

Skrzypacz, A. and H. Hopenhayn (2004): "Tacit collusion in repeated auctions," *Journal of Economic Theory*, 114, 153–169.

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.