Moral Hazard with Counterfeit Signals

Andrew Clausen *
University of Edinburgh

Job market version: November 16, 2011
This version: February 7, 2014

Abstract

In many moral hazard problems, the principal evaluates the agent's performance based on signals which the agent may suppress and replace with counterfeits. This form of fraud may affect the design of optimal contracts drastically, leading to complete market failure in extreme cases. I show that in optimal contracts, the principal deters all fraud, and does so by two complementary mechanisms. First, the principal punishes signals that are suspicious, i.e. appear counterfeit. Second, the principal is lenient on bad signals that the agent could suppress, but does not.

1 Introduction

In high profile cases of corporate earnings manipulation and public medical insurance fraud, the social cost of fraud is quite visible.[1] However, fraud is possible but conspicuously absent in many other markets. For example, websites may defraud advertisers by

*I would like to thank my dissertation committee Steven Matthews, Guido Menzio, and Andrew Postlewaite at the University of Pennsylvania, in addition to Garth Baughman, Efraim Berkovich, Danielle Catambay, David Dillenberger, Roger Dingledine, Suelette Dreyfus, Rachel Greenstadt, Daniel Gottlieb, Aaron Hedland, Kartik Hosanagar, Xiangting Hu, Philipp Kircher, Tzuohann Law, Wei Li, Tony Maida, George Mailath, Kurt Mitman, Nirav Mehta, Marco Ottaviani, Mallesh Pai, Daniel Quint, Iyad Rahwan, Seth Richards, Pierre-André Savalle, Karl Schlag, Carlo Strub, Jonathan Thomas, Xi Weng, and seminar participants at the Australian National Univerisity, Bocconi, Bonn, Edinburgh, McGill, Pompeu Fabra, Royal Holloway, Sciences Po, Vienna, and Yeshiva, for very helpful discussions and support. I may be reached at andrew.clausen@ed.ac.uk.

[1] A string of accounting scandals including Adelphia, Enron, and Worldcom led to the Sarbanes-Oxley Act of 2002. Several large Medicare fraud rings have been uncovered recently; see for example "Justice Department charges 91 in $295 million Medicare fraud scheme" available at http://news.blogs.cnn.com/2011/09/07/justice-department-charges-91-in-295-million-medicare-fraud-scheme/.

inserting and billing for fake clicks on advertisements. Unemployed workers may circumvent government incentives by organizing fake job interviews.[2] Security firms may cover up break-ins that occur on their watch. These types of fraud only occur rarely. But this does not imply that fraud is innocuous in these cases, as the mere possibility of fraud may impose severe hidden costs. For example, consider a moral hazard problem in which the agent may either exert effort, or may use a costless fraud technology that allows him to mimic effort perfectly. Regardless of the contract that the principal offers, the agent prefers to mimic exerting effort. Anticipating this, the principal does not offer any contract to the agent. Therefore there is a complete market failure even though no fraud is committed. More generally, incentives may be distorted substantially by the agent being able to commit fraud. The question of the paper is, how does the possibility of fraud affect the design of incentives in moral hazard problems?

In my model, a risk-neutral principal and a risk-averse agent face a dynamic moral hazard problem. The agent's effort choice in the first period is unobserved, but it determines the distribution of signals realized in the subsequent periods. However, these signals are also not observed by the principal. Rather, the agent is able to suppress some types of signal, and replace them with a counterfeit signal drawn from an exogenous counterfeit signal distribution. The principal pays the agent each period, based on the signals she has observed to date. The agent has a dynamic programming problem to determine which signals to suppress conditional on the history of signal realizations. The principal has an optimal contract design problem, to choose the optimal payment policy to implement a target effort and fraud policy.

In classic moral hazard models, the agent only has one choice to make, so gradual draws of signals may be equivalently modeled as simultaneous draws. However, my model is necessarily dynamic as the agent must decide whether to suppress a signal before observing the next one.

The first result characterizes the fraud policies in optimal contracts. Fraud may be costly for the agent to commit, and may be risky for the agent if there is a chance of getting caught. The first result establishes that "no fraud" is an optimal fraud policy, and that no optimal fraud policy involves any risky or costly fraud; in these cases, the principal may simulate fraud more efficiently than the agent can commit it. To simulate fraud, the principal draws a simulated counterfeit signal from the counterfeit signal distribution, and pays the agent as if she had observed the simulated signal. In other words, the principal may simulate fraud by paying the agent a lottery. If fraud is costly, the principal deducts the costs from the payments, and hence recovers these

[2] In countries with time-unlimited unemployment insurance (such as Australia and the United Kingdom), claimants are required to document their job search activities to continue receiving payments.

costs. If fraud is costless, the principal may improve on the lottery: she may recover the risk premium of the lottery by paying the agent the certainty equivalent. The resulting payment policy deters fraud if it does not improve the agent's ability to distort signals.[3]

The second result characterizes optimal payment policies to deter fraud. As in standard moral hazard models, each signal contains news about the agent's effort. There are two additional attributes of signals. First, if a signal is outside the support of the counterfeit signal distribution, then it is unsuspicious. On the other hand, a signal assigned a higher probability under this distribution is more suspicious, ceteris paribus. Second, a signal may be suppressible or unsuppressible. The second result answers the question: how do these three attributes of signals affect payments? Measures of how good the news is, and how suspicious the news is, are constructed from the signal separately. This means that better news is rewarded more, and more suspicious news is rewarded less, ceteris paribus. However, the payment policy is lenient on bad news that is suppressible. This dampens the incentive to suppress bad news. The lenience of the principal after sufficiently bad suppressible news involves disregarding the severity of the news; therefore optimal contracts may be incomplete.

Finally, the paper explores two special cases. The counterfeit good news case corresponds to the click fraud and fake job interview examples. In this case, either good news or no news arrives every period. If no news arrives, then the agent may fabricate good news. Thus, the principal must measure how suspicious the good news appears, and punish suspicious news adequately to deter fraud. This becomes more costly for her as the real and counterfeit signal distributions become more alike. I argue that the contracts employed in the internet advertising industry match the predictions of the model.

The bad news suppression case corresponds to the example of covering up break-ins, and also to covering up safety and environmental disasters. In this case, either bad news or no news arrives. Bad news that is more costly to suppress is punished more than bad news that is cheaper to suppress. I argue that inefficient incentives in the internet computer security industry led a company, DigiNotar to attempt to cover up an intrusion, ultimately leading to its bankruptcy.

Fraud and incentives for truth-telling have long been a consideration in mechanism design, and there are several strands of related literature. The revelation principle is a central part in the analysis of mechanism design problems in which communication is costless and misrepresentation is undetectable. In these problems, the revelation principle holds very generally and plays a purely technical role to simplify mechanism

[3] The condition used throughout most of the paper is that all signals in the support of the counterfeit signal distribution are unsuppressible. More general conditions are discussed in Section 4.3.

design problems by focusing on direct (non-fraud) mechanisms. However, when the message the agent may send is partially constrained by the state of the world, the issue of misrepresentation becomes more subtle, and the revelation principle only holds under more restricted conditions. Green and Laffont (1986) introduced partially verifiable information to mechanism design, in which the set of messages that the agent is able to send is state-dependent and is a subset of the agent's type space. If evidence satisfies the "nested range condition", then a social choice function is implementable if and only if it is truthfully implementable. Bull and Watson (2007) drop the requirement that the message space is a subset of the type space. They show that if evidence satisfies a weaker condition called "normality", then a strong revelation principle holds in the sense that every mechanism is equivalent to one with full evidence disclosure. Kartik and Tercieux (2011) study full implementation (of Nash equilibrium) where evidence may be fabricated by the agents at some state-dependent cost. They show that for three or more players, a social choice function is implementable without costly fabrication if and only if evidence is "cost-monotonic" and the social choice rule satisfies "no veto power." They do not study whether it is (second-best) optimal to deter costly fabrication and do not accommodate the possibility of a fabrication attempt being unsuccessful.

Townsend (1979) studies a model in which agents may hide their endowments, but the principal may audit them at some cost. In this *costly state verification* model, optimal contracts deter misrepresentation by the use of random audits. While auditing is not explicitly present in my model, the risk of an audit is similar to the risk of a counterfeit signal appearing suspicious. However, in contrast to costly state verification models, the principal never learns for sure if a counterfeit signal is indeed counterfeit.

Costly state falsification was introduced by Lacker and Weinberg (1989), and further studied by Maggi and Rodríguez-Clare (1995), Crocker and Morgan (1998), Crocker and Slemrod (2007) and Crocker and Gresik (2010). My focus is quite different from this literature in two respects. First, this literature assumes that the agent has access to a fraud technology that gives him perfect (but costly) control of signals. I focus on stochastic fraud technologies in which the agent has imperfect control of realized signals. This allows me to study how incentives should respond to noisy signals of fraudulent activity. Second, in this literature, the principal is ignorant about the agent's marginal cost of committing fraud.[4] This ignorance typically implies that fraud occurs in optimal contracts. However, fraud is not ubiquitous, and my focus is on the institutional response to the possibility of fraud, even in markets were fraud is absent. In my model,

[4] Specifically, they assume that at any history, the principal only observes the sum of the state of nature and the fraud committed. The marginal cost of committing more fraud depends on the (unobserved) amount of fraud already committed.

the principal knows what resources are required to conduct fraud at every history, which implies that fraud does not occur in optimal contracts. Nevertheless, the possibility of fraud drastically alters incentives.

Finally, Allen and Gale (1992) considers fraud as a possible answer to the puzzle, why are contracts simpler in practice than in theory? They argue that if an agent is able to manipulate some signals, then optimal contracts are incomplete in the sense that payments are insensitive to those signals. A similar form of incompleteness appears in optimal contracts in my counterfeit signals model as well: all sufficiently bad suppressible news is treated the same. On the other hand, contracts are more complicated in the sense that payments depend on how suspicious the signals are.

The remainder of this paper is organized as follows. Section 2 introduces an example that illustrates the ideas of suspicious information and lenience. Section 3 describes a dynamic moral hazard model with effort and fraud. Section 4 establishes that optimal contracts deter risky and costly fraud. Section 5 applies the first-order approach to characterize the optimal payment policy. Section 6 and Section 7 study special cases of the model, and relate them to click fraud and cover-ups of security intrusions, respectively. Appendix A provides omitted proofs, and Appendix B establishes the validity of the first-order approach for a special case of the model.

2 Example

A risk neutral principal and a risk averse agent would like to trade. The agent has a utility function $u : \mathbb{R} \to \mathbb{R}$ and an outside option w. He chooses effort $e \in \{0, 1\}$, where $e = 1$ means high effort. He receives a private signal $\hat{\theta} \in \{A, B, C\}$ that is distributed

$$\pi(\cdot|e) = \begin{cases} (1, 0, 0) & \text{if } e = 0, \\ (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) & \text{if } e = 1. \end{cases}$$

The A signal is bad news that indicates low effort, and the B and C signals are good news that indicate high effort. The agent has access to a costless fraud technology that allows him to suppress A and replace it with a counterfeit B. I write $f = 1$ if he decides to use this technology, and $f = 0$ otherwise. The resulting signal θ is public. The $\theta = B$ signal may be interpreted as suspicious, because the principal can not distinguish between real and counterfeit B signals. The principal's problem is to choose payments $p(\theta)$ for each state θ in order to minimize the expected cost of implementing effort $e = 1$ subject to a voluntary participation (VP) and an incentive

compatibility (IC) constraint:

$$\begin{aligned}
&\min_{p(\cdot),f} E[p(\theta)|e=1,f] \\
&\text{s.t.} \quad \text{(VP)} \quad E[u(p(\theta))|e=1,f]-1 \geq u(w), \\
&\qquad \text{(IC)} \quad (1,f) \in \underset{(\hat{e},\hat{f})}{\operatorname{argmax}} E[u(p(\theta))|\hat{e},\hat{f}]-\hat{e}.
\end{aligned}$$

This problem has a solution that involves no fraud (i.e. $f=0$). To see this, suppose that $p(\cdot)$ implements $(e,f)=(1,1)$. Then the following payment policy implements $(e,f)=(1,0)$ by simulating fraud:

$$\hat{p}(\theta)=\begin{cases} p(B) & \text{if } \theta=A \text{ or } \theta=B, \\ p(C) & \text{if } \theta=C. \end{cases}$$

Under the payment policy $\hat{p}(\cdot)$, the agent is indifferent between committing fraud or not, since the principal pays him the same amount after A and B signals. The principal's implementation cost is the same under both payment policies, so the new payment policy is also a solution the principal's problem.

Since there is a solution that deters fraud, we may focus on contracts that deter fraud. This simplifies the principal's problem:

$$\begin{aligned}
&\min_{p(\cdot)} \frac{1}{3}[p(A)+p(B)+p(C)] \\
&\text{s.t.} \quad \text{(VP)} \quad \frac{1}{3}[u(p(A))+u(p(B))+u(p(C))]-1 \geq u(w), \\
&\qquad \text{(IC-e)} \quad \frac{1}{3}[u(p(A))+u(p(B))+u(p(C))]-1 \geq u(p(A)), \\
&\qquad \text{(IC-f)} \quad u(p(A)) \geq u(p(B)),
\end{aligned}$$

where (IC-e) requires that effort $e=1$ is optimal and (IC-f) requires that abstaining from fraud is optimal.

It is straightforward to show that if $p^*(\cdot)$ is a solution to the principal's problem, then all three constraints hold with equality at $p^*(\cdot)$. Since all of the constraints share four common expressions, all four of these expressions are equal, so that $w=p^*(A)=p^*(B)<p^*(C)$. The agent is awarded the most after a C signal, as it is good news that is not the result of counterfeiting. Even though B is equally as good news as C, it is rewarded less than C because it is more suspicious. Finally, A is rewarded the same as B, even though it is worse news than B; the principal must be lenient after A to deter fraud.

This optimal contract is not unique; there is also an optimal contract that involves fraud. If the principal pays the agent

$$p'(\theta) = \begin{cases} p^*(A) - \varepsilon & \text{if } \theta = A, \\ p^*(\theta) & \text{if } \theta \neq A, \end{cases}$$

where $\varepsilon > 0$, then the agent will commit fraud and receive the same payment as under $p^*(\cdot)$. Thus, $p'(\cdot)$ implements $(e, f) = (1, 1)$ at the same cost to the principal.

This example illustrates several ideas that are important in the general model. The principal may deter fraud by simulating fraud on the agent's behalf — in the example, by replacing A with B. This means the implementation problem may be simplified by focusing on no-fraud contracts. The optimal contract involves rewarding good news more than bad news ($p(C) > p(A)$), unsuspicious news more than suspicious news ($p(C) > p(B)$), and being lenient on suppressible bad news ($p(A) = p(B)$).

3 Model

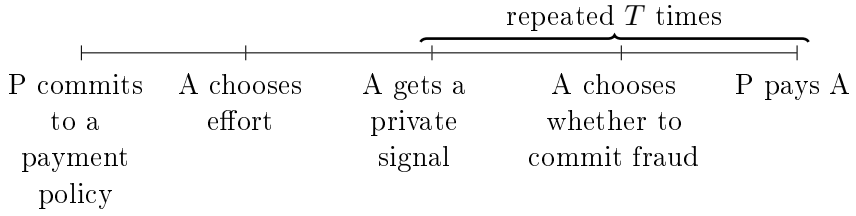

Figure 1: The moral hazard problem with counterfeit signals.

This paper studies a dynamic moral hazard model, which is summarized in Figure 1. There is a risk neutral principal and a risk averse agent with an increasing utility function $u : \mathbb{R} \to \mathbb{R}$. They both discount at rate β over $T + 1$ time periods. In the first period, the agent exerts unobservable effort $e \in [0, 1]$, which the principal values at $v(e)$. In each subsequent time period t, the agent receives a private signal $\hat{\theta}_t$ that is distributed according to $\pi(\cdot|e)$, which has full support over a finite set Θ. The agent has access to a fraud technology, which consists of two parts. The agent first *suppresses* the signal $\hat{\theta}_t$, which costs $c(\hat{\theta}_t) \in \{0, \infty\}$.[5] Then the agent *counterfeits* a replacement signal which is drawn from the *counterfeit signal distribution* $\phi(\cdot)$. The principal observes the resulting public signal θ_t. I assume that the support of $\phi(\cdot)$ only includes *unsuppressible*

[5] I also study an extension with general suppression costs $c(\theta) \in [0, \infty]$ in Section 4.2 and Section 7.

signals, i.e. $c(\theta) = \infty$. This implies that the agent may not conduct fraud more than once in each time period.[6]

At the start of the game, the principal commits to a payment policy $\{p_t(\cdot)\}$. This means that the principal promises to pay the agent $p_t(\theta^t)$ at the end of period t based on the history of public signals $\theta^t = \theta_1, \ldots, \theta_t$. At any time period $t \leq T$, there are two relevant value functions for the agent: the agent's value before the private signal is drawn, $W_t(e; \theta^{t-1})$, and after the private signal is drawn, $V_t(e; \theta^{t-1}, \hat{\theta}_t)$. The agent's history of private signals $\hat{\theta}^{t-1}$ is not payoff relevant, so it is dropped from his Bellman equations:

$$W_t(e; \theta^{t-1}) = \sum_{\hat{\theta}} \pi(\hat{\theta}|e) V_t(e; \theta^{t-1}, \hat{\theta}) \tag{1}$$

$$V_t(e; \theta^{t-1}, \hat{\theta}) = \max \begin{cases} u(p_t(\theta^{t-1}, \hat{\theta})) + \beta W_{t+1}(e; \theta^{t-1}, \hat{\theta}), \\ \sum_{\theta} \phi(\theta) \left[u(p_t(\theta^{t-1}, \theta) - c(\hat{\theta})) + \beta W_{t+1}(e; \theta^{t-1}, \theta) \right], \end{cases} \tag{2}$$

where

$$W_{T+1}(e; \theta^t) = 0.$$

A candidate solution to the agent's dynamic fraud problem in (2) is a fraud policy $f_t : \Theta^t \to \{0, 1\}$, where $f_t(\theta^{t-1}, \hat{\theta}_t) = 1$ means the agent suppresses the signal $\hat{\theta}_t$ after a history of public signals θ^{t-1}. (In the contracting problem below, it will be possible to think of the effort and fraud policy choices being made simultaneously, so e is omitted from the agent's private history in the fraud policy.) The no-fraud policy $\{f_t^*(\cdot)\}_{t=1}^T$ involves no fraud at every history. The agent chooses effort e to maximize $\beta W_1(e) - e$.

This paper studies the principal's implementation problem.

Problem 1. *The principal's implementation problem is to minimize the expected cost of implementing effort e,*

$$\begin{aligned} C(e) = \min_{\{p_t(\cdot)\}, \{f_t(\cdot)\}} & E\left[\sum_{t=1}^{T} \beta^t p_t(\theta^t) \middle| e, \{f_t(\cdot)\} \right] \\ s.t. \quad & \textit{(VP)} \quad \beta W_1(e) - e \geq u_0 \\ & \textit{(IC-e)} \quad e \in \operatorname*{argmax}_{\hat{e} \in [0,1]} \beta W_1(\hat{e}) - \hat{e} \\ & \textit{(IC-f)} \quad f_t(\cdot) \textit{ solves the dynamic programming problem in (1) and (2),} \end{aligned}$$

where the effort e and the fraud policy determines the distribution of θ_t: $\hat{\theta}_t \sim \pi(\cdot|e)$ is

[6] I discuss the role of this assumption and several possible generalizations in Section 4.3.

drawn first, and if $f_t(\theta^{t-1}, \hat{\theta}_t) = 0$ *then* $\theta_t = \hat{\theta}_t$*; otherwise* $\theta_t \sim \phi(\cdot)$*. I say that the payment policy* $\{p_t(\cdot)\}$ *implements effort* e *and fraud policy* $\{f_t(\cdot)\}$ *if these items satisfy the (VP), (IC-*e*) and (IC-*f*) constraints.*

This model has two novel features. Firstly, the agent does not (necessarily) have complete control over the outcome of the fraud. In both the costly state verification and costly state falsification literatures, agents may counterfeit any signal they like, and the fraud technology is entirely predictable. In contrast, this model accommodates unpredictable fraud technologies. For example, if a publisher does a fake click, it does not know what information the advertiser has previously stored about the fake visitor to the website, and can not fully anticipate how suspicious the click will appear. If an agent offers to bribe a witness to suppress information, the agent does not know for sure whether the bribe will be accepted and honored.

Secondly, the agent's fraud decision is dynamic. The agent may base his decisions to commit fraud based on the success or failure of previous attempts at fraud; this allows him to avoid accruing many suspicious signals. Conversely, the agent may not defer fraud decisions; he may not retroactively suppress information that is already public. For example, advertisers observe clicks as they happen. A website publisher may reduce the amount of click fraud it commits based on how suspicious the previous clicks appeared. However, the publisher may not retroactively add fake clicks.

4 Optimal Fraud Policy

Fraud is socially undesirable, as it is unproductive, and potentially includes several social costs. Firstly, fraud may be risky: if there is a chance of being caught, perpetrators take on risks beyond their control. Secondly, Section 4.2 studies an extension in which fraud involves paying a suppression cost. Finally, fraud destroys information, and therefore hampers the provision of incentives. Given that fraud is inefficient, an important question is what level of fraud is optimal, after taking into account the difficulties of incentives. This section shows that under general conditions, the principal may transform contracts with risky fraud into better contracts (with lower implementation costs) without fraud. The logic resembles the revelation principle. Suppose the principal knows the distribution of counterfeit signals and which signals are suppressible. Then the principal may simulate fraud by paying the agent a lottery based on a random draw from the counterfeit signal distribution. Under quite general conditions, the fraud simulation lottery deters fraud. In addition, the principal may replace the lottery draws with their certainty equivalents without disrupting incentives. This allows

the principal to recover the risk premium of fraud.

Section 4.2 studies an extension in which suppressible signals may have a non-zero suppression cost. In this case, the suppression cost is deducted from the fraud simulation lottery. This allows the principal to recover the suppression cost that would be otherwise wasted if the agent actually committed fraud.

The conditions under which optimal contracts deter fraud are quite general. Fraud may only be imperfectly detectable, with the principal never having the possibility of receiving a conclusive signal that it occurred. Fraud may be dynamic. For instance, the agent is free to choose to conduct small amounts of fraud over a long period of time, including that the possibility that an agent could "lie low for a while" after incriminating information comes out. There may or may not be a suppression cost involved. However, none of these issues have any bearing on whether the no-fraud policy is optimal. The key assumption is that the support of the counterfeit signal distribution only contains unsuppressible signals. The role of this assumption and other assumptions are discussed in Section 4.3.

4.1 Suboptimality of Risky Fraud

The following theorem establishes that contracts can always be transformed to eliminate fraud without reducing payoffs. Moreover, if fraud is risky, then the principal may recover the risk premium of fraud.

Definition 1. A contract $(\{p_t(\cdot)\}, e, \{f_t(\cdot)\})$ involves *risky fraud* if at some history $(\theta^{t-1}, \hat{\theta}_t)$,

1. the agent conducts fraud, i.e. $f_t(\theta^{t-1}, \hat{\theta}_t) = 1$, and
2. two possible counterfeit signals θ_t and θ'_t in the support of $\phi(\cdot)$ lead to differing payments at some subsequent time $\tau \geq t$, i.e. $p_\tau(\theta^{t-1}, \theta_t, \cdot) \neq p_\tau(\theta^{t-1}, \theta'_t, \cdot)$.

Theorem 1. *Suppose $\{p_t(\cdot)\}$ implements effort e and fraud policy $\{f_t(\cdot)\}$ in Problem 1. Then there exists a payment policy $\{\hat{p}_t(\cdot)\}$ that implements effort e and the no-fraud policy $\{f^*_t(\cdot)\}$ such that the implementation cost is*

1. *lower if the original contract involves risky fraud, and*
2. *the same otherwise.*

The proof involves the following sequence of transformations. In Lemma 1, a new contract is constructed in which fraud conducted by the agent is delegated to the principal. The principal simulates counterfeit signals with public lotteries drawn every period. The transformation from the original contract to the new contract is reminiscent

of the revelation principle in that the principal simulates the agent's strategy to arrive at a simpler contract.

Lemma 2 shows that the public nature of the lottery draws in the new contract is irrelevant to the agent's incentives. This is because the agent has a weakly dominant strategy, namely no fraud, that is independent of any information he receives throughout the contract. This allows a simpler contract in which the lotteries and payments are hidden from the agent.

In Lemma 3, the principal increases her payoff by replacing the lotteries with their certainty equivalents. This transformation allows the principal to extract the risk premia of fraud. The transformation is possible if fraud is costless; as in the Holmstrom (1979) sufficient statistic theorem, the agent's incentives are preserved throughout this transformation. Therefore, even if fraud is costless to conduct, it is still suboptimal when it imposes gratuitous risk on the agent.

When the agent delegates fraud to the principal in Lemma 1, the principal pays the agent a lottery. As Problem 1 does not include the possibility of paying lotteries, I generalize the notation here. Henceforth, tildes above letters will indicate that a variable is related to a lottery. The principal pays the agent $\tilde{p}_t(\theta^t, \varepsilon^t)$ that depends on the public random draws ε^t. Each ε_t is drawn from a distribution $\psi_t(\cdot|\theta^t; \varepsilon^{t-1})$ at the end of period t. The agent's fraud policy $\{f_t(\cdot;\cdot)\}$ in this setting is a function of both the signals $(\theta^{t-1}, \hat{\theta}_t)$ and the lottery draws ε^{t-1}. The Bellman equations (1) and (2) for the agent generalize to the lottery setting in a natural way:

$$W_t(e; \theta^{t-1}, \varepsilon^{t-1}) = \sum_{\hat{\theta}} \pi(\hat{\theta}|e) V_t(e; \theta^{t-1}, \hat{\theta}; \varepsilon^{t-1}) \tag{3}$$

$$\begin{aligned} &V_t(e; \theta^{t-1}, \hat{\theta}; \varepsilon^{t-1}) \\ &\quad = \max \begin{cases} E_{\varepsilon_t}\left(u(\tilde{p}_t(\theta^{t-1}, \hat{\theta}; \varepsilon^t)) + \beta W_{t+1}(e; \theta^{t-1}, \hat{\theta}; \varepsilon^t)\right), \\ E_{\varepsilon_t}\left(\sum_\theta \phi(\theta)\left[u(\tilde{p}_t(\theta^{t-1}, \theta; \varepsilon^t) - c(\hat{\theta})) + \beta W_{t+1}(e; \theta^{t-1}, \theta; \varepsilon^t)\right]\right). \end{cases} \end{aligned} \tag{4}$$

Fraud involves replacing a signal with a counterfeit drawn from the distribution $\phi(\cdot)$. The principal may simulate fraud by ignoring an observed signal, and paying the agent as if she observed an independent draw from the counterfeit signal distribution. The principal may simulate fraud policy $\{f_t(\cdot)\}$ by

1. drawing ε_t from $\phi(\cdot)$ at histories where the agent would have committed fraud, and
2. setting $\varepsilon_t = \theta_t$ at histories where the agent would not have committed fraud.

This idea leads to the following definition of the fraud simulation lottery.

Definition 2. Given a contract $(\{p_t(\cdot)\}, e, \{f_t(\cdot)\})$ the corresponding *fraud simulation lottery* is

$$\tilde{s}_t(\theta^t; \varepsilon^t) = p_t(\varepsilon^t),$$

where ε_t is drawn from the distribution

$$\psi_t(\varepsilon_t|\theta^t, \varepsilon^{t-1}) = \begin{cases} I(\varepsilon_t = \theta_t) & \text{if } f_t(\varepsilon^{t-1}, \theta_t) = 0, \\ \phi(\varepsilon_t) & \text{if } f_t(\varepsilon^{t-1}, \theta_t) = 1, \end{cases}$$

and $I(\square)$ is the indicator function that is 1 if the proposition $\square$ is true, and 0 otherwise.

In the fraud simulation lottery, ε^t plays the role of the (simulated) public history, and θ^t plays the role of the private history. Even though the principal observes both the simulated public and private histories, she commits to ignoring the private history in order to deter fraud.

Lemma 1. *Suppose the payment policy $\{p_t(\cdot)\}$ implements effort e with fraud policy $\{f_t(\cdot)\}$. Then the fraud simulation lottery $\{\tilde{s}_t(\cdot;\cdot)\}$ implements e with the non-fraud policy $\{f_t^*(\cdot;\cdot)\}$. The implementation cost is the same under both contracts.*

Proof. By construction, the signals in the old contract, θ^t, and the new contract, ε^t, are identically distributed, provided the agent does not deviate from the prescribed no-fraud policy. Hence, the payments are identically distributed in the two contracts.

It remains to show that the new contract respects the fraud incentive constraint (IC-f). Recall that the support of the counterfeit signal distribution only includes unsuppressible signals. This means that regardless of the agent's choices, the fraud simulation lottery ensures that a real or simulated (but not both) counterfeit signal is drawn at every history in which fraud is committed under the original contract, i.e. at every history in which $f_t(\varepsilon^{t-1}, \hat{\theta}_t) = 1$. The only question is whether the agent might profitably deviate by committing fraud at some additional history. But this would contradict the original fraud policy $\{f_t(\cdot)\}$ being optimal under the original payment rule. Hence the no-fraud policy $\{f_t^*(\cdot;\cdot)\}$ is optimal under the new contract. $\square$

Lemma 1 showed that fraud may be replaced with lotteries without disrupting the agent's incentives. The rest of the proof of Theorem 1 involves showing that if these lotteries are replaced with their certainty equivalents, the principal may recover the risk premia of fraud without disrupting incentives. However, the notion of certainty equivalent is subtle in this dynamic setting. The payment in time period t depends

on the entire history of lottery draws ε^t, so that uncertainty about ε_t creates risk for the agent in future periods. Therefore, the right notion of certainty equivalent must be applied to capture the dynamic risk of the dynamic lotteries and to ensure the dynamic no-fraud incentive constraints are respected.

The following lemma addresses relationship between the lottery draws ε^t and the dynamic incentives. Lemma 2 establishes that the lottery draws and the resulting payments may be hidden from the agent, since the no-fraud action is a weakly dominant at every history for him. This *hidden lottery setting* is purely of technical interest, as the agent does not observe his own utility. In this setting, the lottery draws are no longer a state variable, and the agent's Bellman equation (4) becomes

$$\begin{aligned} &V_t(e;\theta^{t-1},\hat{\theta}) \\ &= \max \begin{cases} E_{\varepsilon^t}\left(u(\tilde{p}_t(\theta^{t-1},\hat{\theta};\varepsilon^t))\right) + \beta W_{t+1}(e;\theta^{t-1},\hat{\theta}), \\ \sum_\theta \phi(\theta)\left[E_{\varepsilon^t}\left(u(\tilde{p}_t(\theta^{t-1},\theta;\varepsilon^t) - c(\hat{\theta}))\right) + \beta W_{t+1}(e;\theta^{t-1},\theta)\right]. \end{cases} \end{aligned} \tag{5}$$

Lemma 2. *If a lottery payment policy $\{\tilde{p}_t(\cdot,\cdot)\}$ implements e with the no-fraud policy in Problem 1, then it implements e with the no-fraud policy in the hidden lottery setting of (5).*

Proof. Since $\{\tilde{p}_t(\cdot,\cdot)\}$ implements the no-fraud policy, the agent has a weakly dominant strategy. This means the realizations of ε^t are irrelevant to the agent's decisions. □

Finally, Lemma 3 shows that the lotteries used to simulate counterfeit signals can be replaced by their certainty equivalents. This does not affect the agent's payoffs, but the principal's payoff is increased by the risk premium of the lottery. Therefore, the principal is better off deterring fraud whenever it is risky for the agent.

Lemma 3. *Suppose $\tilde{p}_t(\cdot;\cdot)$ is a lottery payment policy that implements effort e and the no-fraud policy $\{f_t^*(\cdot)\}$) in the hidden lotteries setting. Then the certainty equivalent payment policy defined by*

$$u(\hat{p}_t(\theta^t)) = E_{\varepsilon^t}[u(\tilde{p}_t(\theta^t,\varepsilon^t))|\theta^t] \tag{6}$$

also implements $(e,\{f_t^*(\cdot)\})$. It is less costly for the principal whenever any of the payments $\tilde{p}_t(\theta^t;\cdot)$ are non-degenerate lotteries.

Proof. Since $c(\hat{\theta}) \in \{0,\infty\}$ for all $\hat{\theta} \in \Theta$, it follows that fraud deviation payoffs are

preserved by the new payment policy, i.e.

$$u(\hat{p}_t(\theta^t) - c(\hat{\theta})) = E_{\varepsilon^t}[u(\tilde{p}_t(\theta^t, \varepsilon^t) - c(\hat{\theta}))|\theta^t]. \tag{7}$$

Substituting (6) and (7) into the agent's Bellman equation (2) in the non-lottery setting gives the Bellman equation:

$$\begin{aligned} &V_t(e; \theta^{t-1}, \hat{\theta}) \\ &= \max \begin{cases} E_{\varepsilon^t}\left[E_{\varepsilon^t}[u(\tilde{p}_t(\theta^t, \varepsilon^t))|\theta^t]\right] + \beta W_{t+1}(e; \theta^{t-1}, \hat{\theta}), \\ \sum_{\theta_t} \phi(\theta_t) \left(E_{\varepsilon^t}\left[E_{\varepsilon^t}[u(\tilde{p}_t(\theta^t, \varepsilon^t) - c(\hat{\theta}))|\theta^t]\right] + \beta W_{t+1}(e; \theta^t)\right). \end{cases} \end{aligned} \tag{8}$$

After removing the redundant expectations (by the law of iterated expectations), this becomes the value function (5) under the lottery payment policy $\{\tilde{p}_t(\cdot;\cdot)\}$. Since the value functions are equivalent under the two payment policies, by the Principle of Optimality, they implement the same actions.

If any of the lotteries $\tilde{p}_t(\theta^t;\cdot)$ are non-degenerate, then the new payment policy is less costly to the principal. The principal's expected payment is lowered by the risk premium of this lottery, which is greater than zero. This follows from a standard argument by Holmstrom (1979). By Jensen's inequality, the concavity of u implies

$$u(p_t(\theta^t)) = E_{\varepsilon^t}[u(\tilde{p}_t(\theta^t; \varepsilon^t))|\theta^t] < u\left(E_{\varepsilon^t}[\tilde{p}_t(\theta^t; \varepsilon^t)|\theta^t]\right). \tag{9}$$

Since u is increasing it follows that

$$p_t(\theta^t) < E_{\varepsilon^t}[\tilde{p}_t(\theta^t; \varepsilon^t)|\theta^t]. \square$$

This completes the proof of Theorem 1, that optimal contracts deter risky fraud. In fact, Lemma 2 and Lemma 3 also establish that lotteries are suboptimal:

Corollary 1. *If the no-fraud contract* $(\{\tilde{p}_t(\cdot;\cdot)\}, e, \{f_t^*(\cdot,\cdot)\})$ *is a solution to the lottery extension of Problem 1 of implementing effort* e*, then* $\{\tilde{p}_t(\cdot;\cdot)\}$ *is a degenerate lottery.*

4.2 Extension: Suboptimality of Costly Fraud

The analysis above established that optimal contracts deter risky fraud, because the principal may simulate fraud more efficiently than the agent may commit it. Another reason that fraud may be inefficient is that the agent may have to pay a suppression cost $c(\hat{\theta}) \in (0, \infty)$ to suppress the signal $\hat{\theta}$. This section shows optimal contracts deter costly fraud, because the principal may recover the suppression cost with an appropriate

fraud simulation payment policy. The fraud simulation lottery may be generalized to this setting in a straightforward way: the principal deducts the suppression cost of fraud from payments at the histories where it simulates fraud for the agent.

In the costly fraud setting, the suppression cost is any function $c : \Theta \to [0, \infty]$. An important difference in this setting is that Lemma 3 no longer holds, so that lottery payment policies may be a feature of optimal contracts, and the principal may not be able to recover the risk premium of risky fraud. To see this, suppose the agent has decreasing absolute risk aversion. Then if the agent pays a suppression cost, his risk aversion would increase, so lotteries may be used by the principal to increase the effective cost of fraud to the agent. Consequently, this section expands the set of possible payment policies the principal may choose to include lotteries, as defined in (4). On the other hand, if the agent has constant absolute risk aversion, then Lemma 3 holds in the costly fraud setting.

The fraud simulation lottery in Definition 2 generalizes to the costly fraud setting as follows. Given a contract $(\{\tilde{p}_t(\cdot;\cdot)\}, e, \{f_t(\cdot;\cdot)\})$ with lottery draws $\varepsilon_t \sim \psi_t(\cdot|\theta^t, \varepsilon^{t-1})$, the corresponding fraud simulation lottery is

$$\tilde{s}_t(\theta^t; \varepsilon^t, \delta^t) = \tilde{p}_t(\delta^t; \varepsilon^t) - f_t(\delta^{t-1}, \theta_t; \varepsilon^{t-1}) c(\theta_t),$$

where $(\varepsilon_t, \delta_t)$ is drawn from

$$\rho_t(\varepsilon_t, \delta_t | \theta^t; \varepsilon^{t-1}, \delta^{t-1}) = \psi_t(\varepsilon_t | \theta^t; \varepsilon^{t-1}) \begin{cases} I(\delta_t = \theta_t) & \text{if } f_t(\delta^{t-1}, \theta_t; \varepsilon^{t-1}) = 0, \\ \phi(\delta_t) & \text{if } f_t(\delta^{t-1}, \theta_t; \varepsilon^{t-1}) = 1. \end{cases}$$

Definition 3. The contract $(\{\tilde{p}_t(\cdot;\cdot)\}, e, \{f_t(\cdot;\cdot)\})$ involves *costly fraud* if at some history $(\theta^{t-1}, \hat{\theta}_t; \varepsilon^{t-1})$,

1. the agent conducts fraud, i.e. $f_t(\theta^{t-1}, \hat{\theta}_t; \varepsilon^{t-1}) = 1$, and
2. the signal $\hat{\theta}_t$ has a strictly positive suppression cost, i.e. $c(\hat{\theta}_t) > 0$.

Proposition 1. *If the lottery $\{\tilde{p}_t(\cdot,\cdot)\}$ implements effort e and fraud policy $\{f_t(\cdot,\cdot)\}$, then the fraud simulation lottery $\{\tilde{s}_t(\cdot,\cdot)\}$ implements effort e and the no-fraud policy $f_t^*(\cdot,\cdot)$. The implementation cost under the fraud simulation lottery is*

1. *lower if the original contract involves costly fraud, and*
2. *the same otherwise.*

Proof. The proof is analogous to Lemma 1. □

4.3 Robustness of Optimal Fraud Policy

The results above establish general conditions under which optimal contracts deter fraud. This section explores the assumptions that lead to this result, and how they may break down.

Counterfeit signals are unsuppressible

One key assumption is that all signals in the support of the counterfeit distribution are unsuppressible. This assumption implies that the agent may only commit fraud once each time period. I illustrate with an example that if the assumption is dropped, but the restriction of one fraud attempt per period is retained, then the no-fraud results of Theorem 1 and Proposition 1 do not hold. This is because the fraud simulation lottery effectively allows the agent to circumvent the limit of fraud attempts. However, if both the assumption and the one-attempt restriction are dropped, then these no-fraud results hold.

Suppose $T = 1$ and that there are three possible private signals, $\hat{\theta} \in \{A, B, C\}$. After effort $e = 0$, the signal is drawn uniformly from $\{A, B\}$, and after effort $e = 1$, the signal $\hat{\theta}$ is uniformly distributed over all possible signals. The fraud technology allows A to be suppressed and replaced with a uniform draw from $\{A, B\}$. Unlike Problem 1, the counterfeit signal distribution includes a suppressible signal A in its support.

Suppose that the agent may not attempt to suppress A more than once. The fraud simulation lottery in this setting is $\tilde{s}(\theta, \varepsilon) = p(\varepsilon)$, where ε is drawn uniformly from $\{A, B\}$ if $\theta = A$, and $\varepsilon = \theta$ if $\theta \neq A$. Lemma 1 fails in this example: the fraud simulation lottery does not deter fraud. If the agent commits fraud under this payment policy, then the probability of either the agent or the principal replacing A with B is $\frac{3}{4}$; the agent is effectively given two chances to suppress the bad news A. This means that when the agent is limited to one suppression attempt, the fraud simulation effectively expands the set of possible fraud strategies available to him. This problem is avoided when the assumption that all signals in the support of the counterfeit distribution are unsuppressible is made: the fraud simulation lottery does not introduce any new possibilities for the agent in this case.

Another way to avoid this problem is to allow the agent to repeatedly suppress signals until he draws an unsuppressible signal (B). In this case, the fraud simulation lottery does not allow the agent to conduct any more fraud than he could under the original contract. Thus, the main assumption needed for Lemma 1 to hold is that the fraud simulation policy does not expand the feasible set of fraud policies that the agent

may choose from. If this assumption is met, then the fraud simulation lottery deters fraud, since the principal simulates an optimal fraud policy for the agent.

There are other ways an asymmetry may arise when the principal pays the agent a fraud simulation lottery. For example, suppose the agent may repeatedly suppress signals, but with an increasing marginal suppression cost each time. Then at any given history, the principal does not know the cost the agent faces for committing fraud (since she does not know how much fraud was committed in the past). Therefore, the principal can not faithfully simulate fraud on behalf of the agent. This explains why fraud is optimal in the model of Lacker and Weinberg (1989), which has an increasing marginal cost of falsification.

Multiple Fraud Technologies

In Problem 1, the agent only has access to a single fraud technology. However, this is not important for the results. Suppose that the agent had access to a set of fraud technologies Y, where each technology $y \in Y$ consists of a suppression cost function $c_y(\cdot)$ and a counterfeit signal distribution $\phi_y(\cdot)$. Theorem 1 generalizes to this setting. In particular, the principal may simulate fraud by drawing from the appropriate counterfeit signal distribution at each history (and subtracting the appropriate suppression cost, as in the extension in Section 4.2).

Unknown Fraud Technology

If the principal is not fully informed about the fraud technology, then all optimal contracts may involve fraud. The fraud simulation lottery idea does not apply directly, as the principal does not know ex ante which counterfeit signal distribution(s) are available to the agent, and at what cost(s). It may be feasible for the principal to induce the agent to reveal its private information about the fraud technology to the principal. However, this is not optimal in general. For example, if the agent has a sufficiently low probability of having access to a fraud technology, the cost from distorting the payment policy to screen the agent types would outweigh the benefit of recovering the risk premia and/or suppression costs of fraud.

5 Optimal Payment Policy

The previous section showed that under general conditions, optimal contracts deter risky and costly fraud, and can be adapted to deter all fraud. However, the question remains: what are the optimal incentives to deter fraud, and how do they interact with

incentives for productive effort? Signals vary in the extent to which they are good news about the agent's effort, whether or not they are suppressible, and if they are suppressible, how suspicious they are. Thus, there are three aspects of signals that are potentially relevant for payments. This section establishes that the principal separately evaluates the good news and suspicious news aspects. Better news is paid more, and suspicious news is paid less. However, this is not the end of the story. The principal is lenient on suppressible bad news, which reduces the incentives for the agent to commit fraud. Equally bad news that is unsuppressible is not forgiven in this way, because harsh punishments after unsuppressible signals do not invite fraud (as it is infeasible for the agent). Thus, all three dimensions of the signals are relevant for payments: better news is rewarded more, suspicious news is rewarded less, and suppressible bad news is treated leniently compared to equally bad unsuppressible news.

This section begins by reformulating the principal's problem in a manner suitable for taking first-order conditions. This involves writing the constraints in a sequence form (rather than a dynamic programming form), and applying the first-order approach to simplify the agent's effort constraint. Then the section proceeds to use first-order conditions to characterize the payment policy.

5.1 Problem Reformulation

The goal of this section is to transform Problem 1 into a form in which first-order conditions may be applied to characterize the optimal payment policy. There are two approaches in dynamic contract theory: Rogerson (1985b) applies a variational approach based on the sequence problem, and Spear and Srivastava (1987) and Thomas and Worrall (1990) simplify the problem by applying a dynamic programming approach with promised utility as the state variable.

While dynamic programming was helpful in studying the agent's problem, its application to the principal's problem is problematic. Firstly, the model has persistence, in the sense that the effort choice e affects information and behavior in every subsequent period. This means that an additional state variable (promised marginal benefit of effort) would be required, and the agent's incentive constraints would need to be reformulated in terms of both state variables.[7] Secondly, different fraud incentive constraints bind at different histories, and at states where the binding constraint changes, the principal's value function would be non-differentiable.

In contrast to the difficulties of the dynamic programming approach, the varia-

[7] This is similar to the analysis of Williams (2011), in which the "promised marginal utility of the private state" is an additional state variable in the principal's dynamic programming problem.

tional approach of Rogerson (1985b) is relatively simple. In the repeated moral hazard problem of Rogerson (1985b), optimal contracts may involve the agent playing a non-stationary strategy. However, in view of Theorem 1, optimal contracts in the counterfeit signals model involve an effort choice followed by the stationary no-fraud strategy. To summarize, the single persistent effort choice, followed by a sequence of identical non-fraud choices adds minimal complexity to the principal's implementation problem. Therefore, this section reformulates the principal's problem using a variational approach rather than a dynamic programming approach.

The rest of the paper focuses on Problem 2, which involves implementing the no-fraud policy. Theorem 1 established that the implementation cost is the same as in Problem 1, even though the principal is more constrained.

Problem 2. *The principal's no-fraud implementation problem is to minimize the expected cost of implementing effort e,*

$$
\begin{aligned}
C(e) = \min_{\{p_t(\cdot)\}} \; & E\left[\sum_{t=1}^{T} \beta^t p_t(\theta^t) \middle| e\right] \\
\text{s.t.} \quad & \textit{(VP)} \qquad \beta W_1(e) - e \geq u_0 \\
& \textit{(IC-}e\textit{)} \qquad e \in \operatorname*{argmax}_{\hat{e}\in[0,1]} \beta W_1(\hat{e}) - \hat{e} \\
& \textit{(IC-}f^*\textit{)} \quad f_t^*(\cdot) \textit{ solves the dynamic programming problem in (1) and (2),}
\end{aligned}
$$

where $\theta_t \sim \pi(\cdot|e)$.

The rest of the section reformulates the (VP), (IC-e) and (IC-f) constraints into a form suitable for taking first-order conditions. Since the agent finds the no-fraud policy is optimal at effort e, the (VP) constraint may be rewritten non-recursively as

$$
\text{(VP}') \qquad\qquad E\left[\sum_{t=1}^{T} \beta^t u(p_t(\theta^t)) \middle| e\right] - e \geq u_0.
$$

The no-fraud constraint, (IC-f^*) may also be rewritten non-recursively. By the one-shot deviation principle, the constraint may be formulated as a set of constraints, one for each history $\hat{\theta}^t$, that requires the agent to prefer the no-fraud continuation policy over suppressing $\hat{\theta}_t$ only. Adopting the convention that all expectations are with respect

to the real signal distribution $\pi(\cdot|e)$, the constraint may be rewritten as

$$\begin{aligned}(\text{IC-}f^*) \quad \text{for all } \hat{\theta}^\tau, \quad & E\left[\sum_{t=\tau}^{T} \beta^t u(p_t(\theta^t)) \middle| e, \theta^\tau = \hat{\theta}^\tau\right] \\ & \geq E\left[\frac{\phi(\theta_\tau)}{\pi(\theta_\tau|e)}\left(\beta^\tau u(p_\tau(\theta^\tau) - c(\hat{\theta}_\tau)) + \sum_{t=\tau+1}^{T} \beta^t u(p_t(\theta^t))\right) \middle| e, \theta^{\tau-1} = \hat{\theta}^{\tau-1}\right].\end{aligned}$$

Note that the expectation for counterfeiting includes the likelihood ratio, $\frac{\phi(\theta_\tau)}{\pi(\theta_\tau|e)}$ in order to replace the (implicitly included) real signal distribution with the counterfeit signal distribution.

Finally, the (IC-e) constraint is complicated because it involves a continuum of inequalities (one for each e), each of which is complex because the agent may deviate to a different fraud policy for different effort levels $\hat{e} \neq e$. A common simplification is to replace (IC-e) with its first-order condition,

$$W_1'(e) = 1.$$

This is problematic for two reasons. Firstly, first-order conditions are merely necessary (not sufficient) for the agent to find e an optimal effort choice. Thus, this new constraint involves a relaxation of the principal's problem. Appendix B establishes that the relaxed and original problems share the same solution when there is one time period ($T = 1$), under some mild conditions on the signal distribution π. It is an open question whether the first-order approach is valid when there are more time periods.

Secondly, the first-order condition is not well-defined, as the value function W_1 is not differentiable at all effort levels. The agent's value function is the upper envelope of a set of differentiable functions, one for each fraud policy. The upper envelope need not be differentiable at effort choices where the agent is indifferent between two fraud policies. However, the constraint only requires W_1 to be differentiable at the implemented effort e. Theorem 1 of Clausen and Strub (2011) shows that the value function is differentiable at e, since e is an optimal choice for the agent. Moreover, the same theorem establishes that $W_1'(e)$ may be evaluated at the agent's optimal fraud

policy (i.e. no-fraud). Hence,

$$W_1'(e) = \left[\frac{d}{d\hat{e}} E\left[\sum_{t=1}^{T} \beta^t u(p_t(\theta^t)) \middle| \hat{e}\right]\right]\Bigg|_{\hat{e}=e} \tag{10}$$

$$= E\left[\sum_{t=1}^{T} \frac{\pi_e(\theta^t|e)}{\pi(\theta^t|e)} \beta^t u(p_t(\theta^t) \middle| e\right] \tag{11}$$

$$= E\left[\sum_{t=1}^{T}\sum_{\tau=1}^{t} \frac{\pi_e(\theta_\tau|e)}{\pi(\theta_\tau|e)} \beta^t u(p_t(\theta^t) \middle| e\right]. \tag{12}$$

This establishes that

$$\text{(IC-}e'\text{)} \qquad E\left[\sum_{t=1}^{T}\sum_{\tau=1}^{t} \frac{\pi_e(\theta_\tau|e)}{\pi(\theta_\tau|e)} \beta^t u(p_t(\theta^t)) \middle| e\right] = 1.$$

is a relaxation of (IC-e).

After replacing the (VP), (IC-e) and (IC-f^*) constraints with the constraints above, the reformulated problem is suitable for taking first-order conditions:

Problem 3. *The principal's reformulated implementation problem is*

$$\begin{aligned}
C(e) = &\min_{p_t(\cdot)} E\left[\sum_{t=1}^{T} \beta^t p_t(\theta^t) \middle| e\right] \\
&s.t. \quad \textit{(VP')} \quad E\left[\sum_{t=1}^{T} \beta^t u(p_t(\theta^t)) \middle| e\right] - e \geq u_0 \\
&\qquad \textit{(IC-}e'\textit{)} \quad E\left[\sum_{t=1}^{T}\sum_{\tau=1}^{t} \frac{\pi_e(\theta_\tau|e)}{\pi(\theta_\tau|e)} \beta^t u(p_t(\theta^t)) \middle| e\right] = 1. \\
&\qquad \textit{(IC-}f^*\textit{)} \quad \textit{for all } \hat{\theta}^\tau, \quad E\left[\sum_{t=\tau}^{T} \beta^t u(p_t(\theta^t)) \middle| e, \theta^\tau = \hat{\theta}^\tau\right] \\
&\qquad\qquad \geq E\left[\frac{\phi(\theta_\tau)}{\pi(\theta_\tau|e)} \left(\beta^\tau u(p_\tau(\theta^\tau) - c(\hat{\theta}_\tau)) + \sum_{t=\tau+1}^{T} \beta^t u(p_t(\theta^t))\right) \right. \\
&\qquad\qquad\qquad \left. \middle| e, \theta^{\tau-1} = \hat{\theta}^{\tau-1}\right],
\end{aligned}$$

where all expectations are taken with respect to $\theta_t \sim \pi(\cdot|e)$.

5.2 Characterization

Let $(\lambda, \mu, \{\nu_t(\theta^t)\})$ be the Lagrange multipliers on the voluntary participation constraint, effort incentive constraint, and no-fraud incentive constraint at history θ^t. The following theorem characterizes the optimal payment policy in the two cases, that the fraud constraint is slack (i.e. (IC-f^*) holds with strict inequality) or binds (i.e. (IC-f^*) holds with equality).

Theorem 2. *If $\{p_t^*(\cdot)\}$ is an optimal solution to Problem 3 with Lagrange multipliers $(\lambda^*, \mu^*, \{\nu_t^*(\cdot)\})$, then:*

1. *If the no-fraud incentive constraint (IC-f^*) is slack at history θ^t, then*

$$\frac{1}{u'(p_t^*(\theta^t))} = \underbrace{\frac{1}{u'(p_{t-1}^*(\theta^{t-1}))}}_{\text{yesterday}} + \mu^* \underbrace{\frac{\pi_e(\theta_t|e)}{\pi(\theta_t|e)}}_{\substack{\text{good news} \\ \text{measure}}} - \bar{\nu}_t^*(\theta^{t-1}) \underbrace{\frac{\phi(\theta_t)}{\pi(\theta_t|e)}}_{\substack{\text{suspicious news} \\ \text{measure}}}, \tag{13}$$

   where the first term on the right is replaced by λ^* for $t = 1$, and

$$\bar{\nu}_t^*(\theta^{t-1}) = \sum_{\hat{\theta}_t} \nu_t^*(\theta^{t-1}, \hat{\theta}_t). \tag{14}$$

2. *If the no-fraud incentive constraint (IC-f^*) binds at histories (θ^{t-1}, θ_t) and $(\theta^{t-1}, \theta_t')$, then $p_\tau^*(\theta^{t-1}, \theta_t, \cdot) = p_\tau^*(\theta^{t-1}, \theta_t', \cdot)$ for all $\tau \geq t$.*

The first part identifies three components of payments when the no-fraud constraint does not bind, i.e. at histories where the agent strictly prefers not to destroy the signal and replace it with a counterfeit. The terms in the first-order condition may be interpreted as follows. In a pure insurance problem (absent any moral hazard or fraud concerns), the principal's first-order condition would be

$$\frac{1}{u'(p_t^*(\theta^t))} = \underbrace{\frac{1}{u'(p_{t-1}^*(\theta^{t-1}))}}_{\text{yesterday}},$$

which means that the agent's payment is the same at every history, i.e. complete insurance and consumption smoothing. Adding in moral hazard incentive concerns (but leaving aside fraud concerns), the principal's first-order condition would be

$$\frac{1}{u'(p_t^*(\theta^t))} = \underbrace{\frac{1}{u'(p_{t-1}^*(\theta^{t-1}))}}_{\text{yesterday}} + \mu^* \underbrace{\frac{\pi_e(\theta_t|e)}{\pi(\theta_t|e)}}_{\substack{\text{good news} \\ \text{measure}}},$$

which means that the agent is rewarded more after better news about its effort e. The marginal likelihood ratio in the last term is a standard measure of good news in moral hazard problems, and is discussed by Milgrom (1981). Briefly, higher values represent better news about the agent's effort, positive and negative values are possible, and its expected value is zero. Finally, in the moral hazard with counterfeit signals model, the first-order condition is

$$\frac{1}{u'(p_t^*(\theta^t))} = \underbrace{\frac{1}{u'(p_{t-1}^*(\theta^{t-1}))}}_{\text{yesterday}} + \mu^* \underbrace{\frac{\pi_e(\theta_t|e)}{\pi(\theta_t|e)}}_{\substack{\text{good news} \\ \text{measure}}} - \bar{\nu}_t^*(\theta^{t-1}) \underbrace{\frac{\phi(\theta_t)}{\pi(\theta_t|e)}}_{\substack{\text{suspicious news} \\ \text{measure}}},$$

which means the agent is rewarded less after more suspicious news. In the last term, the measure of suspiciousness is the likelihood ratio that the signal was drawn from the counterfeit distribution versus the real distribution (for the contracted effort level e). Higher likelihood ratios indicate more suspicious signals. The likelihood ratio is multiplied by $\bar{\nu}_t^*(\theta^{t-1})$, which is the sum of the Lagrange multipliers at that history. The $\nu_t(\theta^{t-1}, \hat{\theta}_t)$ multiplier is non-zero only if the no-fraud incentive constraint binds for $\hat{\theta}_t$. These multipliers are summed because deterring fraud affects payments in the same way, regardless of the signal being suppressed: signals that appear suspicious are punished.

The first-order condition (13) only applies if the no-fraud constraint is slack at θ^t. If it binds, then the first-order condition would pay the agent too little, which would give the agent an incentive to suppress the signal. In other words, if the signal is unsuppressible, the principal is at liberty to punish the agent according to the first-order condition. If the signal is suppressible, then the principal must be lenient after sufficiently bad news. When the principal is lenient in this way, the agent is paid enough to be indifferent between committing fraud or not. If the no-fraud constraint binds for two histories (θ^{t-1}, θ_t) and $(\theta^{t-1}, \theta_t')$, then the agent's expected discounted utility is the same, since the payoff from fraud is the same in both cases. The second part of the theorem establishes that these histories are treated identically at all subsequent time periods. This means that if θ_t is worse news than θ_t', then θ_t is forgiven more. The principal finds it optimal to discard sufficiently bad news from consideration — but only if the news is suppressible. This result is similar to a finding of Allen and Gale (1992). In their setting, fraud gives rise to incomplete contracts that disregard informative signals.

The proof of the first part involves elementary manipulation of the first-order conditions of Problem 3, so it is in the appendix. The proof of the second part appears below. The proof strategy is to suppose (for the sake of contradiction) that the payments do

differ across the two signals θ_t and θ'_t, and construct a new payment policy that stochastically swaps the two signals. The new payment policy implements the same effort and the no-fraud policy at the same implementation cost — but as in Corollary 1, the principal could reduce its implementation cost be replacing the lottery with its certainty equivalent. This is a contradiction.

Proof of Theorem 2 part 2. Suppose for the sake of contradiction that at some histories $(\eta^{\tau-1}, \eta_\tau)$ and $(\eta^{\tau-1}, \eta'_\tau)$, the no-fraud constraint binds but for some $\tau' \geq \tau$, $p^*_{\tau'}(\eta^{\tau-1}, \eta_\tau, \cdot) \neq p^*_{\tau'}(\eta^{\tau-1}, \eta'_\tau, \cdot)$. This proof will construct a new payment policy that implements the same effort e and the no-fraud policy at lower implementation cost.

Consider the payment policy $\{\hat{p}_t(\cdot)\}$ that swaps $(\eta^{\tau-1}, \eta_\tau)$ and $(\eta^{\tau-1}, \eta'_\tau)$, i.e.

$$\hat{p}_t(\theta^t) = \begin{cases} p^*_t(\eta^{\tau-1}, \eta'_\tau, \theta_{\tau+1}, \ldots, \theta_t) & \text{if } \theta^\tau = (\eta^{\tau-1}, \eta_\tau), \\ p^*_t(\eta^{\tau-1}, \eta_\tau, \theta_{\tau+1}, \ldots, \theta_t) & \text{if } \theta^\tau = (\eta^{\tau-1}, \eta'_\tau), \\ p^*_t(\theta^t) & \text{otherwise.} \end{cases}$$

Since the no-fraud constraint binds at both histories, the agent's value is the same, i.e.

$$W_{\tau+1}(e; \eta^{\tau-1}, \eta_\tau) = W_{\tau+1}(e; \eta^{\tau-1}, \eta'_\tau).$$

Under the new payment policy, the agent's value $\{\hat{W}_t(e; \cdot)\}$ is obtained by swapping $(\eta^{\tau-1}, \eta_\tau)$ and $(\eta^{\tau-1}, \eta'_\tau)$ in $\{W_t(e; \cdot)\}$. Before time τ, the value functions under both payment policies are the same, so fraud is deterred by the principal of optimality. After time τ, the value functions are swapped between the two histories, but fraud is deterred by both continuation payment policies, so fraud is deterred under the new payment policy $\{\hat{p}_t(\cdot)\}$.

However, the new payment policy has a different probability distribution of payments (at effort e). This change may be undone by randomizing between the two payment policies based on a random draw $\varepsilon_\tau \in \{\eta_\tau, \eta'_t\}$ at the histories (η^{t-1}, η_t) and (η^{t-1}, η'_t). (The random draws at all other histories are degenerate.) For example, one such lottery payment policy is

$$\tilde{p}_t(\theta^t; \varepsilon^t) = \begin{cases} p^*_t(\theta^t) & \text{if } \theta^{\tau-1} = \eta^{\tau-1},\ \theta_\tau \in \{\eta_\tau, \eta'_\tau\}, \text{ and } \varepsilon_\tau = \theta_\tau, \\ \hat{p}_t(\theta^t) & \text{if } \theta^{\tau-1} = \eta^{\tau-1},\ \theta_\tau \in \{\eta_\tau, \eta'_\tau\}, \text{ and } \varepsilon_\tau \neq \theta_\tau, \\ p^*_t(\theta^t) & \text{otherwise,} \end{cases}$$

where ε_τ is drawn from

$$\psi_\tau(\varepsilon_\tau|\theta^\tau, \varepsilon^{\tau-1}) = \pi(\varepsilon_\tau|e, \varepsilon_\tau \in \{\eta_\tau, \eta'_\tau\}).$$

Since both the original payment policy $\{p_t^*(\cdot)\}$ and the swapped payment policy $\{\hat{p}_t(\cdot)\}$ deter fraud, the lottery payment policy $\{\tilde{p}_t(\cdot;\cdot)\}$ that randomizes between them deters fraud as well. Since the probability distribution of payments is the same under $\{\tilde{p}_t(\cdot;\cdot)\}$ and $\{p_t^*(\cdot)\}$ when the agent follows the no-fraud policy, the left side of the effort incentive constraint (IC-e') is the same in both cases. This is most simply seen by studying the form of the effort constraint based on (11):

$$E_{\theta^T, \varepsilon^T}\left[\sum_{t=1}^{T} \frac{\pi_e(\theta^t|e)}{\pi(\theta^t|e)} \beta^t u(\tilde{p}_t(\theta^t; \varepsilon^t)) \middle| e\right] = E_{\theta^T}\left[\sum_{t=1}^{T} \frac{\pi_e(\theta^t|e)}{\pi(\theta^t|e)} \beta^t u(p_t^*(\theta^t)) \middle| e\right].$$

Thus, $\{\tilde{p}_t(\cdot;\cdot)\}$ implements effort e and the no-fraud policy in Problem 3 at the same implementation cost as $\{p_t^*(\cdot)\}$. However, lottery payment policies are suboptimal, as Corollary 1 generalizes to the setting of Problem 3. Thus, $\{p_t^*(\cdot)\}$ is a suboptimal payment policy — a contradiction. □

To summarize, optimal payment policies in the moral hazard with counterfeit signals model are closely related to those in standard moral hazard settings. As in standard settings, the payment policies feature insurance and consumption smoothing as well as higher payments after better news about the agent's effort. However, to deter fraud, optimal payments reward suspicious signals less, and are lenient on suppressible bad news.

6 Counterfeit Good News

In the click fraud and fake job interview examples, the agent is able to replace "no news" with counterfeit good news. While all good news signals are equally indicative of high effort, some good news signals appear more suspicious than others. For example, a click originating from an IP address within the organized computer crime group, the Russian Business Network (discussed below), is suspicious good news. Similarly, an unemployed worker's claim to have attended a job interview with a close relative is suspicious good news. The principal's problem is to reward good news without creating perverse incentives for creating fake good news. This section the studies counterfeit good news case, and finds that the implementation cost increases as the real and counterfeit good news signals become more alike. This is followed by an argument that the model matches

Google's controversial history with click fraud.

6.1 Model

In the *counterfeit good news case* of Problem 3, the agent's effort increases the arrival probability $a(e)$ of good news, which is an unsuppressible signal $\hat{\theta}_t$ drawn from some distribution $\pi_G(\cdot)$. Otherwise no news arrives, which is represented by the null signal $\hat{\theta}_t = 0$. Thus, the private signals are drawn from

$$\pi(\hat{\theta}|e) = \begin{cases} 1 - a(e) & \text{if } \hat{\theta} = 0, \\ a(e)\pi_G(\hat{\theta}) & \text{if } \hat{\theta} \neq 0. \end{cases}$$

The agent may suppress the null signal at no cost, in which case a counterfeit signal is drawn from $\phi(\cdot)$. In the counterfeit good news setting, $\text{Supp}(\phi) \subseteq \text{Supp}(\pi_G) = \Theta \setminus \{0\}$, so it is possible to study what happens as these distributions become more alike.

6.2 Analysis

The first-order condition (13) from Theorem 2 implies that after good news $\theta_t \neq 0$, the agent is paid

$$\frac{1}{u'(p_t^*(\theta^t))} = \underbrace{\frac{1}{u'(p_{t-1}^*(\theta^{t-1}))}}_{\text{yesterday}} + \mu^* \underbrace{\frac{a'(e)}{a(e)}}_{\substack{\text{good news} \\ \text{measure}}} - \bar{\nu}_t(\theta^{t-1}) \underbrace{\frac{\phi(\theta_t)}{\pi(\theta_t|e)}}_{\substack{\text{suspicious news} \\ \text{measure}}}. \tag{15}$$

As usual, more suspicious news is rewarded less. In general, it is not possible to rank unsuspicious bad news versus suspicious good news. But in this case, it is clear that to deter fraud, at least one signal has to be paid less than the null signal – otherwise fraud would be a dominant strategy. Therefore, sufficiently suspicious news is paid less than no news.

Proposition 2. *If $\{p_t(\cdot)\}$ is a solution to the counterfeit good news case of Problem 3, then some (suspicious) good news signal is rewarded less than no news. That is, at every history θ^{t-1}, there is some signal $\underline{\theta} \neq 0$ such that*

$$u(p_t(\theta^{t-1}, \underline{\theta})) + \beta W_{t+1}(e; \theta^{t-1}, \underline{\theta}) \leq u(p_t(\theta^{t-1}, 0)) + \beta W_{t+1}(e; \theta^{t-1}, 0). \tag{16}$$

Proof. If (16) were violated for every $\underline{\theta} \neq 0$, then multiplying both sides by $\phi(\underline{\theta})$ and

summing up would gives

$$\sum_{\underline{\theta}\neq 0}\phi(\underline{\theta})\left[u(p_t(\theta^{t-1},\underline{\theta}))+\beta W_{t+1}(e;\theta^{t-1},\underline{\theta})\right] > u(p_t(\theta^{t-1},0))+\beta W_{t+1}(e;\theta^{t-1},0).$$

This means the agent would strictly prefer to commit fraud — a contradiction. □

The following proposition shows that the principal's implementation problem becomes weakly more difficult as the real and counterfeit signal distributions, $\pi_G(\cdot)$ and $\phi(\cdot)$ become more alike. Let $C(e;\alpha)$ be the implementation cost of effort e when the counterfeit signals are distributed according to $\phi' = \alpha\phi + (1-\alpha)\pi_G$, where $\alpha \in [0,1]$. As α decreases, ϕ' and π_G become more alike.

Proposition 3. *In the counterfeit good news setting, $C(e;\cdot)$ is weakly decreasing and $C(e;0)=\infty$.*

The proof strategy is as follows: if a payment policy implements $(e,\{f_t^*(\cdot)\})$ under the counterfeit signal distribution ϕ', then it also implements $(e,\{f_t^*(\cdot)\})$ under the counterfeit signal distribution $\phi(\cdot)$. The key step to establishing this is to imagine allowing the agent to make draws of counterfeit signals from the real distribution rather than the counterfeit distribution, keeping the payment policy fixed. The agent would always prefer draws from the real distribution. If this were not the case, then the (absurd) payment policy that throws out all good news signals, and replaces them with a good news signal drawn from the real distribution, would also deter fraud.

Lemma 4. *If $\{p_t(\cdot)\}$ is a solution to the counterfeit good news case of Problem 3, then the agent would prefer a (hypothetical) counterfeit signal draw from $\pi_G(\cdot)$ rather than $\phi(\cdot)$ at every history $\eta^{\tau-1}$, i.e.*

$$\begin{aligned} &E\left[\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right] \\ &\leq E\left[\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right]. \end{aligned} \tag{17}$$

Proof. Suppose for the sake of contradiction that (17) is violated at some history $\eta^{\tau-1}$: The rest of the proof constructs a new lottery-based payment policy that implements e at the same implementation cost, in violation of Corollary 1. Consider the new lottery payment policy

$$\tilde{p}_t(\theta^t;\varepsilon^t) = p(\varepsilon^t)$$

defined in terms of the following distribution. Let ε_t be the random variable that equals θ_t, except at a history $(\eta^{\tau-1}, \eta_\tau)$ in which $\eta_\tau \neq 0$ is good news, in which case it is a random draw from $\pi_G(\cdot)$,

$$\psi_t(\varepsilon_t|\theta^t;\varepsilon^{t-1}) = \begin{cases} I(\varepsilon = \theta_t) & \text{if } \varepsilon^{t-1} \neq \eta^{\tau-1} \text{ or } \theta_t = 0, \\ \pi_G(\varepsilon_t) & \text{if } \varepsilon^{t-1} = \eta^{\tau-1} \text{ and } \theta_t \neq 0. \end{cases} \tag{18}$$

Under the no-fraud policy, $\{\tilde{p}_t(\cdot;\cdot)\}$ has the same probability distribution of payments as $\{p_t(\cdot)\}$ and implements e. Moreover, since the agent strictly prefers draws from $\phi(\cdot)$ rather than $\pi_G(\cdot)$, the agent prefers not to conduct fraud at history $(\eta^{\tau-1}, 0)$:

$$\begin{aligned}
&E_{\varepsilon^t,\theta^t}\left[\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(\tilde{p}_t(\theta^t;\varepsilon^t))\middle| e, \varepsilon^{\tau-1} = \eta^{\tau-1}, \theta_\tau \neq 0\right] \\
&= E_{\theta^t}\left[\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e, \theta^{\tau-1} = \eta^{\tau-1}, \theta_\tau \neq 0\right] && \text{(by construction)} \\
&< E_{\theta^t}\left[\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e, \theta^{\tau-1} = \eta^{\tau-1}, \theta_\tau \neq 0\right] && \text{(by (17))} \\
&\leq E_{\theta^t}\left[\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e, \theta^{\tau} = (\eta^{\tau-1}, 0)\right] && \text{(since } \{p_t(\cdot)\} \text{ deters fraud)} \\
&= E_{\varepsilon^t,\theta^t}\left[\sum_{t=\tau}^{T}\beta^t u(\tilde{p}_t(\theta^t;\varepsilon^t))\middle| e, (\varepsilon^{\tau-1}, \theta_\tau) = (\eta^{\tau-1}, 0)\right] && \text{(by construction).}
\end{aligned}$$

Therefore, the lottery payment policy $\{\tilde{p}_t(\cdot;\cdot)\}$ implements e optimally, in contradiction to Corollary 1. □

Proof of Proposition 3. Firstly, it is straightforward to verify that any effort $e > 0$ is unimplementable when $\alpha = 0$, i.e. when counterfeit and real signals are identically distributed.

Secondly, fix some $\alpha \in (0,1)$, and set $\phi' = \alpha\phi + (1-\alpha)\pi_G$. Suppose $\{p_t(\cdot)\}$ is an optimal payment policy for implementing e^* given the counterfeit signal distribution ϕ' in Problem 3. Since the counterfeit signal distribution does not appear in the (VP′) and (IC-e') constraints, the contract $(\{p_t(\cdot)\}, e, \{f_t^*(\cdot)\})$ satisfies these constraints under ϕ. By Lemma 4, the agent prefers counterfeit signal draws from π_G than ϕ' at every history

$\eta^{\tau-1}$, i.e.

$$E\left[\left(\alpha\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}+(1-\alpha)\right)\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right] \tag{19}$$
$$\leq E\left[\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right].$$

This implies that the agent prefers draws from π_G over ϕ:

$$E\left[\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right] \tag{20}$$
$$\leq E\left[\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right].$$

Since a draw from ϕ' is a randomization between a draw from π_G and ϕ, it follows that the agent prefers draws from ϕ' over ϕ:

$$E\left[\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right] \tag{21}$$
$$\leq E\left[\frac{\phi'(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right].$$

This establishes that the contract also satisfies the (IC-f^*) constraint under ϕ at every history $(\eta^{\tau-1},0)$:

$$\begin{aligned}
&E\left[\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\middle| e,\theta^{\tau}=(\eta^{\tau-1},0)\right]\\
&\geq E\left[\frac{\phi'(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\ \middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right] \qquad \text{(by (IC-}f^*\text{) under } \phi')\\
&\geq E\left[\frac{\phi(\theta_\tau)}{\pi_G(\theta_\tau)}\sum_{t=\tau}^{T}\beta^t u(p_t(\theta^t))\ \middle| e,\theta^{\tau-1}=\eta^{\tau-1},\theta_\tau\neq 0\right] \qquad \text{(by (21))}.
\end{aligned}$$

Therefore, the contract is also feasible under ϕ, so the implementation cost under counterfeit distribution ϕ is not higher than under ϕ'. □

6.3 Click Fraud

In the internet advertising industry, advertisers pay website publishers to direct visitors to the advertisers' pages. Advertisers prefer to place their ads on high quality websites that attract the most visitors. However, advertisers do not observe website quality. Rather, they only observe the clicks from each website. But this is problematic: publishers may attempt to defraud advertisers with fake clicks.[8] Thus, advertisers face a problem like Problem 1: how should they pay website publishers to create good websites and direct visitors to the ads when clicks may be fabricated?

Google is an intermediary in the internet advertising market that solves the advertiser's problem on the advertiser's behalf. Google (and its competitors) collect information about each click, and decide how much to charge advertisers and reward publishers. When the industry was in its infancy, advertisers were concerned that they were paying for fake clicks. In 2006, Google paid a $90M settlement in a class action lawsuit led by Lane's Gifts that alleged that Google colluded with publishers to defraud advertisers with click fraud. Google attempted to reassure advertisers that they had solved the click fraud problem:

> By far, most invalid clicks are caught by our automatic filters and discarded *before* they reach an advertiser's bill. (3/8/2006)
>
> Undetected click fraud constitute less than 0.02% of all clicks. (2/28/2007)

However, leading computer security pundits were convinced this problem was difficult to solve:

> Google's $6 billion-a-year advertising business is at risk. Google is testing a new advertising model to deal with click fraud: cost-per-action ads. — Bruce Schneier (7/13/2006)
>
> Why Google Click Fraud is NOT 0.02% — `zdnet.com` (3/1/2007)

On the other hand, Google's CEO (but not Google's public relations spokespersons) at the time believed that even if click fraud were rampant, this would not be a problem:[9]

[8] Simple attempts at click fraud, such as repeatedly clicking on ads, are easy to detect by tracking IP (internet protocol) addresses. More sophisticated fraud based on botnets of hacked computers are more difficult to detect. Markets for these botnets are provied by organized crime networks such as the Russian Business Network of St Petersburg (see "Fatal System Error", by Joseph Menn (2010)).

[9] He added that Google polices clickfraud because " we don't like it, and because it does, at least for the short-term, creates some problems before the advertiser sees it, we go ahead and try to detect it and eliminate it."

> Eventually, the price that the advertiser is willing to pay for the conversion will decline, because the advertiser will realize that these are bad clicks, in other words, the value of the ad declines, so over some amount of time, the system is in-fact, self-correcting. In fact, there is a perfect economic solution which is to let it happen. — Eric Schmidt (3/3/2006)

Today, Google's claim that it has click fraud under control is uncontested in the technical press. Google still uses a pay-per-click mechanism (rather than another type such as "cost-per-action"), which discards suspicious clicks and blacklists publishers with too many suspicious clicks. Google still maintains a team of engineers to improve their click fraud detection algorithms.

These observations are consistent with the counterfeit good news model. Theorem 1 asserts that optimal contracts deter risky fraud, which matches Google's claim of deterring almost all fraud rather than Schmidt's view that "a perfect economic solution... is to let it happen." Theorem 2 asserts that payments are lower after more suspicious signals, which matches Google's policy of discarding suspicious clicks. Proposition 2 asserts that a sufficiently suspicious signal is worse than no news, which loosely matches Google's policy of blacklisting websites with too many suspicious clicks. Even though it is optimal to deter fraud regardless of the signal distributions, Proposition 3 shows that there is a value to improving the fraud detection technology, which is consistent with Google's investment in these technologies.

7 Bad News Suppression

In the security firm example, the firm might suppress intrusions as a substitute for preventing them. A similar issue arises in safety and environmental regulation, in which regulators aim to enforce minimum standards on industrial plants. Namely, a plant's management may prefer to suppress incidents rather than follow regulations.[10] This section studies the setting in which the agent may exert effort to decrease the probability of an incident, and may also attempt to suppress news about the incident. Incidents are equally indicative of low effort, but some incidents may appear more suspicious than others, i.e. they may appear like failed suppression attempts. This section shows that incidents with higher suppression costs are treated less leniently in

[10] A familiar example is depicted in the film, *Erin Brockovich*, which according to her is "true and probably 98% accurate." (See `http://www.brockovich.com/movie.html`.) Pacific Gas and Electric attempted to suppress hexavalent chromium contamination in Hinkley, California by buying affected houses. Brockovich was a paralegal assisting with the real estate transactions and became suspicious when she found medical records in the case files.

optimal contracts. This result, as well as the main theorems are discussed in the context of computer security intrusions.

7.1 Model

The *bad news suppression* model is based on the reformulated Problem 3, but with the costly fraud extension described in Section 4.2. For simplicity, this section focuses on the one fraud period case (i.e. $T = 1$). The agent's effort decreases the arrival probability $a(e)$ of bad news, which is a signal distributed according to $\pi_B(\cdot)$. Otherwise no news arrives, which is represented by the null signal $\hat{\theta} = 0$. After bad news $\hat{\theta}$ arrives, the agent may attempt to suppress it at cost $c(\hat{\theta}) \in [0, \infty]$. A new signal θ would then be drawn from $\phi(\cdot)$, which includes no news ($\theta = 0$) and perhaps some unsuppressible bad news signals in its support. As discussed in Section 4.2, lotteries may be optimal in this setting. This leads to the following implementation problem.

Problem 4. *The principal's problem in the bad news suppression case is*

$$
\begin{aligned}
C(e) = & \min_{\tilde{p}(\cdot;\cdot)} E\left[\tilde{p}(\theta,\varepsilon)|e\right] \\
& s.t. \quad \textit{(VP')} \quad E\left[u(\tilde{p}(\theta,\varepsilon))|e\right] - e \geq u_0 \\
& \qquad \textit{(IC-e')} \quad E\left[\frac{\pi_e(\theta|e)}{\pi(\theta|e)} u(\tilde{p}(\theta,\varepsilon)) \middle| e\right] = 1. \\
& \qquad \textit{(IC-f*)} \quad \textit{for all } \hat{\theta} \in \Theta, \\
& \qquad\qquad E_\varepsilon\left[u(\tilde{p}(\hat{\theta},\varepsilon)) \middle| \hat{\theta}\right] \geq E_{\varepsilon,\theta}\left[\frac{\phi(\theta)}{\pi(\theta|e)} u(\tilde{p}(\theta,\varepsilon) - c(\hat{\theta})) \middle| e\right],
\end{aligned}
$$

where all expectations are taken with respect $\theta \sim \pi(\cdot|e)$ and $\varepsilon \sim \psi(\cdot|\theta)$.

7.2 Analysis

The following proposition establishes that if the suppression cost of bad news θ' is higher than that of θ'', then the payment for the signal θ' is weakly lower. This result generalizes part 2 of Theorem 2, which establishes that punishments are limited when suppression is either costless or infinitely costly. This section studies the intermediate case, and shows that punishments are more lenient for signals with lower suppression costs.

Proposition 4. *Suppose $\tilde{p}(\cdot;\cdot)$ is an optimal lottery payment policy in Problem 4.*

1. *If θ is a suppressible signal, then it is awarded a degenerate lottery. That is, there exists $p(\theta)$ such that $p(\theta) = \tilde{p}(\theta;\varepsilon)$ for all ε.*

2. *If θ' and θ'' are suppressible signals with $c(\theta') < c(\theta'')$, then $p(\theta') \geq p(\theta'')$.*

The proof is in the appendix.

7.3 Computer Security Intrusions

On September 20, 2011, the Dutch computer security firm DigiNotar declared bankruptcy.[11] It was a firm entrusted by all of the major web browsers to certify websites belonged to who they claimed.[12] DigiNotar discovered on July 19 that its servers had been hacked, and faced an important choice: should it announce that it had been hacked and suffer an immediate loss of reputation, or should it attempt to suppress the intrusion? If the hackers only planned to exploit the intrusion for small-scale attacks such as breaking into a small number of email accounts, then the intrusion into DigiNotar would probably go unnoticed. If on the other hand the hackers exploited the intrusion for a widespread attack on many targets, then the intrusion would be discovered very publicly, and a major scandal might ensue. DigiNotar decided to suppress the intrusion, and issued a press release the following day claiming that "DigiNotar's certificates are among the most reliable in the field."

Unfortunately for DigiNotar, the attackers were politically motivated and wanted to attract attention. The attackers exploited DigiNotar to issue hundreds of fake certificates for websites including Google, Skype, and Iranian dissident forums. By August 28, three hundred thousand Iranian Google accounts had been hacked, and the general public learned about the security failure. DigiNotar later admitted to covering up the intrusion, but had already been blacklisted by the major web browsers, and bankruptcy was inevitable.

DigiNotar's customers were not the only people affected by this intrusion. Since every major web browser trusted DigiNotar to practice careful security policies, every internet user was potentially affected by the intrusion. For example, even though Google was not in any contractual relationship with DigiNotar, the intrusion allowed Iranian hackers to intercept Google's communications. Therefore, we should think of the principal in this moral hazard problem as being either the web browser developers or regulators (such as the Internet Corporation of Assigned Numbers and Names), who might consider choosing optimal policies to implement adequate security proce-

[11] These events are documented in http://en.wikipedia.org/wiki/DigiNotar, which cites a comprehensive list of news stories. The news report by Charles Arthur, "Rogue web certificate could have been used to attack Iran dissidents" is helpful for describing the political context. It is available at http://www.guardian.co.uk/technology/2011/aug/30/faked-web-certificate-iran-dissidents.

[12] This type of firm is called a *certificate authority*.

dures for internet infrastructure. Since security firms may attempt to suppress intrusion incidents, the principal faces an implementation problem like Problem 4.

In the DigiNotar incident, the suppression was essentially costless (DigiNotar merely chose not to revoke the fake certificates), albeit with a risk of an enormous loss. In this case, optimal contracts deter suppression of intrusion incidents: Theorem 1 implies that in optimal contracts, the agent is deterred from taking on the risks associated with fraud. This suggests that the current arrangements for internet security firms are inefficient, as DigiNotar had an incentive to take on a large risk to suppress the intrusion. Theorem 2 suggests how web browser developers or regulators might improve welfare: optimal contracts make punishments after intrusions occur, and suspicious intrusions that appear like failed suppression attempts are punished more.

In other situations, suppression of bad news is costly. For example, if *qui tam* whistleblower incentives were available to DigiNotar employees, then DigiNotar might attempt to bribe its employees to keep the intrusion secret.[13] In this case, Proposition 1 also implies that optimal contracts deter suppression of intrusion incidents. The general characterization of optimal payment policies of Theorem 2 does not apply. However, Proposition 4 does apply: intrusion incidents that are more costly to suppress are treated less leniently. For example, if the intrusion was in a system visible to many employees, then bribes would be more expensive, so the optimal punishment would be relatively severe.

8 Conclusion

This paper studied a class of moral hazard problems in which, in addition to choosing an unobservable productive effort, the agent has access to a fraud technology that allows him to suppress signals and replace them with counterfeits. This form of fraud is inefficient as it exposes the agent to gratuitous risk, involves unproductive costly activity, and hampers incentive provision. The first main result establishes that every optimal contract may be transformed into another optimal contract without fraud, and optimal contracts do not involve risky or costly fraud. The second main result shows that the principal uses two mechanisms to deter fraud: punishing suspicious signals and being lenient on bad signals that the agent declined the opportunity to suppress. All sufficiently bad suppressible news is treated identically: the optimal payment policy ignores how bad the news is.

[13] In the United States, the False Claims Act, 31 U.S.C. § 3730 allows whistleblowers to file civil suits against contractors that defraud the federal government and receive a "qui tam" portion of the damages.

These results offer an explanation of why the internet advertising market evolved into its current form in which click fraud is deterred, suspicious clicks are discarded, and intermediaries invest in click fraud detection technology. The results also suggest that the large risks taken by internet security firms (specifically, certificate authorities) in suppressing major intrusions are the result of suboptimal incentives.

While the paper focused on how the possibility of fraud affects incentives, it raises two future directions for explaining why fraud occurs. Fraud may occur if the principal is unable to commit to being lenient on the agent, or if the principal is poorly informed about the agent's fraud technology. More generally, this paper provides a benchmark for understanding fraud. If fraud is rampant in some industry such as public medical insurance, one might ask: does the principal punish suspicious news, and is she lenient on suppressible bad news? If either answer is "no", then a potential explanation for the fraud has been identified.

References

Allen, F. and Gale, D. (1992). Measurement distortion and missing contingencies in optimal contracts. *Economic Theory*, **2** (1), 1–26.

Bull, J. and Watson, J. (2007). Hard evidence and mechanism design. *Games and Economic Behavior*, **58** (1), 75–93.

Clausen, A. and Strub, C. (2011). Envelope theorems for non-smooth and non-concave optimization.

Conlon, J. (2009). Two new conditions supporting the first-order approach to multisignal principal-agent problems. *Econometrica*, **77** (1), 249–278.

Crocker, K. J. and Gresik, T. (2010). Optimal compensation with earnings manipulation: Managerial ownership and retention, mimeo.

— and Morgan, J. (1998). Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts. *Journal of Political Economy*, **106** (2), 355–375.

— and Slemrod, J. (2007). The economics of earnings manipulation and managerial compensation. *RAND Journal of Economics*, **38** (3), 698–713.

Green, J. R. and Laffont, J.-J. (1986). Partially verifiable information and mechanism design. *Review of Economic Studies*, **53** (3), 447–456.

Holmstrom, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, **10** (1), 74–91.

Jewitt, I. (1988). Justifying the first-order approach to principal-agent problems. *Econometrica*, **56** (5), 1177–1190.

Kartik, N. and Tercieux, O. (2011). Implementation with evidence: Complete information. *Theoretical Economics*, **Forthcoming**.

Lacker, J. M. and Weinberg, J. A. (1989). Optimal contracts under costly state falsification. *Journal of Political Economy*, **97** (6), 1345–1363.

Maggi, G. and Rodríguez-Clare, A. (1995). Costly distortion of information in agency problems. *RAND Journal of Economics*, **26** (4), 675–689.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, **12** (2), 380–391.

Rogerson, W. P. (1985a). The first-order approach to principal-agent problems. *Econometrica*, **53** (6), 1357–1367.

— (1985b). Repeated moral hazard. *Econometrica*, **53** (1), 69–76.

Spear, S. E. and Srivastava, S. (1987). On repeated moral hazard with discounting. *The Review of Economic Studies*, **54** (4), 599–617.

Thomas, J. and Worrall, T. (1990). Income fluctuation and asymmetric information: An example of a repeated principal-agent problem. *Journal of Economic Theory*, **51** (2), 367–390.

Townsend, R. M. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, **21** (2), 265–293.

Williams, N. (2011). Persistent private information. *Econometrica*, **79** (4), 1233–1275.

A Omitted Proofs

Proof of Theorem 2, claim 1. The first-order condition with respect to $p_t^*(\theta^t)$ in Problem 3 is

$$\begin{aligned}0 =& -\pi(\theta_t|e)\beta^t + \lambda^*\pi(\theta_t|e)\beta^t u'(p_t^*(\theta^t)) + \mu^*\pi(\theta_t|e)\sum_{\tau=1}^{t}\frac{\pi_e(\theta_\tau|e)}{\pi(\theta_\tau|e)}\beta^t u'(p_t^*(\theta^t))\\ &+\sum_{\tau=1}^{t}\nu_\tau^*(\theta^\tau)\beta^t u'(p_t^*(\theta^t)) - \sum_{\tau=1}^{t}\sum_{\hat\theta}\nu_\tau^*(\theta^{\tau-1},\hat\theta)\pi(\theta_t|e)\frac{\phi(\theta_t|e)}{\pi(\theta_t|e)}\beta^t u'(p_t^*(\theta^t))\end{aligned}$$

which can be rewritten as

$$\begin{aligned}1 =&\lambda u'(\theta_t, p_t^*(\theta^t)) + \mu\sum_{\tau=1}^{t}\frac{\pi_e(\theta_\tau|e)}{\pi(\theta_\tau|e)}u'(p_t^*(\theta^t))\\ &+\sum_{\tau=1}^{t}\nu_\tau^*(\theta^\tau)\frac{u'(p_t^*(\theta^t))}{\pi(\theta_t|e)} - \sum_{\tau=1}^{t}\sum_{\hat\theta}\nu_\tau^*(\theta^{\tau-1},\hat\theta)\frac{\phi(\theta_t|e)}{\pi(\theta_t|e)}u'(p_t^*(\theta^t))\end{aligned}$$

and

$$\frac{1}{u'(p_t^*(\theta^t))} = \lambda^* + \mu^*\sum_{\tau=1}^{t}\frac{\pi_e(\theta_\tau|e)}{\pi(\theta_\tau|e)} + \sum_{\tau=1}^{t}\nu_\tau(\theta^\tau)\frac{1}{\pi(\theta_\tau|e)} - \sum_{\tau=1}^{t}\sum_{\hat\theta}\nu_\tau^*(\theta^{\tau-1},\hat\theta)\frac{\phi(\theta_\tau)}{\pi(\theta_\tau|e)}.$$

This can be rewritten recursively as

$$\frac{1}{u'(p_t^*(\theta^t))} = \frac{1}{u'(p_{t-1}^*(\theta^{t-1}))} + \mu^*\frac{\pi_e(\theta_t|e)}{\pi(\theta_t|e)} + \nu_t^*(\theta^t)\frac{1}{\pi(\theta_t|e)} - \frac{\phi(\theta_t)}{\pi(\theta_t|e)}\sum_{\hat\theta}\nu_t^*(\theta^{t-1},\hat\theta), \tag{22}$$

If the no-fraud constraint is slack at history θ^t, then $\nu_t^*(\theta^t) = 0$, which gives expression (13). □

Proof of Proposition 4. Let $(\lambda, \mu, \nu(\cdot))$ be the Lagrange multipliers for the voluntary participation, effort incentive, and fraud incentive constraints. The first-order conditions with respect to $\tilde p(\theta;\varepsilon)$ may be written as

$$\frac{1}{u'(\tilde p(\theta,\varepsilon))} = \lambda + \mu\frac{a'(e)}{a(e)} + \nu(\theta)\frac{1}{\pi(\theta|e)} - \sum_{\hat\theta}\nu(\hat\theta)\frac{u'(\tilde p(\theta,\varepsilon) - c(\hat\theta))}{u'(\tilde p(\theta,\varepsilon))}\frac{\phi(\theta)}{\pi(\theta|e)}. \tag{23}$$

If θ is suppressible, then $\phi(\theta) = 0$, so (23) simplifies to

$$\frac{1}{u'(\tilde{p}(\theta,\varepsilon))} = \lambda + \mu\frac{a'(e)}{a(e)} + \nu(\theta)\frac{1}{\pi(\theta|e)}, \tag{24}$$

which does not depend on ε. This establishes the first part.

If the no-fraud constraints bind at both θ' and θ'', then

$$u(p(\theta')) = E_{\varepsilon,\theta}\left[\frac{\phi(\theta)}{\pi(\theta|e)}u(\tilde{p}(\theta,\varepsilon) - c(\theta'))\middle|e\right],$$
$$u(p(\theta'')) = E_{\varepsilon,\theta}\left[\frac{\phi(\theta)}{\pi(\theta|e)}u(\tilde{p}(\theta,\varepsilon) - c(\theta''))\middle|e\right],$$

which implies that $u(p(\theta')) > u(p(\theta''))$. On the other hand, if the no-fraud constraint does not bind at θ'', then $\nu(\theta'') = 0$ so that

$$\frac{1}{u'(\tilde{p}(\theta,\varepsilon))} = \lambda + \mu\frac{a'(e)}{a(e)}. \tag{25}$$

Since $\nu(\theta') \geq 0$, this establishes $p(\theta') \geq p(\theta'')$. □

B Validity of the First-Order Approach

The analysis in Section 5 studies a first-order approach relaxation of the principal's problem. This section provides conditions under which the first-order approach is valid, i.e. the solutions to the two problems coincide. The techniques of Rogerson (1985a), Jewitt (1988), and Conlon (2009) are inapplicable here. To see this, consider the first-order condition for the one-time period case of the principal's problem,

$$\frac{1}{u'(p(\theta))} = \lambda + \mu\frac{\pi_e(\theta|e)}{\pi(\theta|e)} - \nu\frac{\phi(\theta)}{\pi(\theta|e)}.$$

The approach of the aforementioned papers is to show that the right side is increasing and concave in either θ or e. Their conditions are applicable to the first two terms, but not the final term. In particular, there is no reason to assume that better signals about the agent's effort are less suspicious. (On the contrary, agents prefer counterfeiting technologies that mimic high effort.)

This section establishes that the relaxed problem is equivalent to the original problem under the following conditions:

1. there is only one time period ($T = 1$), and

2. the signal distribution can be decomposed into a convex combination

$$\pi(\theta|e) = a(e)\pi_G(\theta) + (1 - a(e))\pi_B(\theta),$$

of two distributions that do not depend on e, and

3. the arrival probability $a(\cdot)$ in this decomposition is differentiable and concave in effort e with $a'(e) > 0$ for all e.

Note that the distributions in the two special cases (counterfeiting good news and suppressing bad news) are special cases of the second condition.

The first condition implies that if (p^*, e^*, f^*) is an optimal contract in the relaxed problem, then the agent's value function may be written as

$$W(e) = \sum_{\theta \in \Theta} \pi(\theta|e) u(p^*(\theta)).$$

This is because, the agent does not find it optimal to commit fraud even after deviating from effort e^*, as the no-fraud incentive constraint (IC-f^*) does not depend on e in the one period case (nor in the final period in the general case).

Under the second condition, choosing e to maximize $W(e) - e$ is isomorphic to choosing $A = a(e)$ to maximize

$$\sum_{\theta \in \Theta} \left[A\pi_G(\theta) + (1 - A)\pi_B(\theta)\right] u(p^*(\theta)) - a^{-1}(A). \tag{26}$$

That is, e^* maximizes $W(e) - e$ if and only if $A^* = a(e^*)$ maximizes (26).

The third condition implies that (26) is concave. This means that first-order conditions are sufficient for identifying maximizers of (26). By the chain rule and the condition that $a' > 0$, the set of stationary points of $W(e) - e$ and (26) are isomorphic (i.e. $W'(e^*) = 1$ if and only if $A^* = a(e^*)$ is a stationary point of (26)). Therefore, if $W'(e^*) = 1$, then $A^* = a(e^*)$ is a stationary point of (26), and A^* maximizes (26), so e^* maximizes $W(e) - e$.

To summarize, under the three conditions, the first-order condition $W'(e^*) = 1$ is sufficient for establishing that e^* is an optimal choice for the agent. Hence, the optimal solution to the relaxed problem is also feasible (and hence optimal) in the original problem.