ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

ECONOMIC ANALYSIS OF INSURANCE FRAUD

Pierre PICARD

August 2012

Cahier n° 2012-23

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.economie.polytechnique.edu/
mailto:chantal.poujouly@polytechnique.edu

ECONOMIC ANALYSIS OF INSURANCE FRAUD[1]

Pierre PICARD[2]

August 2012

Cahier n° 2012-23

Abstract: We survey recent developments in the economic analysis of insurance fraud. The paper first sets out the two main approaches to insurance fraud that have been developped in the literature, namely the costly state verification and the costly state falsification. Under costly state verification, the insurer can verify claims at some cost. Claims' verification may be deterministic or random, and it can be conditioned on fraud signals perceived by insurers. Under costly state falsification, the policyholder expends resources for the building-up of his or her claim not to be detected. We also consider the effects of adverse selection, in a context where insurers cannot distinguish honest policyholders from potential defrauders, as well as the consequences of credibility constraints on anti-fraud policies. Finally, we focus attention on the risk of collusion between policyholders and insurance agents or service providers.

Keywords: Fraud, audit, verification, falsification, collusion, build-up.

JEL classification: D80, G22

[1] This article is Chapter 10 in the second edition of the *Handbook of Insurance*, Georges Dionne (Ed.), Springer, New York, forthcoming in 2013; ISBN 978-1-4614-0154-4.

[2] Department of Economics, Ecole Polytechnique, Palaiseau, France

Economic Analysis of Insurance Fraud

Pierre Picard

Abstract

We survey recent developments in the economic analysis of insurance fraud. The paper first sets out the two main approaches to insurance fraud that have been developped in the literature, namely the costly state verification and the costly state falsification. Under costly state verification, the insurer can verify claims at some cost. Claims' verification may be deterministic or random, and it can be conditioned on fraud signals perceived by insurers. Under costly state falsification, the policyholder expends resources for the building-up of his or her claim not to be detected. We also consider the effects of adverse selection, in a context where insurers cannot distinguish honest policyholders from potential defrauders, as well as the consequences of credibility constraints on anti-fraud policies. Finally, we focus attention on the risk of collusion between policyholders and insurance agents or service providers.

Keywords: Fraud, audit, verification, falsification, collusion, build-up.

JEL Classification Numbers: D80, G22

1 Introduction

Insurance fraud is a many-sided phenomenon[1]. Firstly, there are many different degrees of severity in insurance fraud, going from build-up to the planned criminal fraud, through opportunistic fraud. Furthermore, insurance fraud refers primarily to the fact that policyholders may misreport the magnitude of their losses[2] or report an accident that never occured, but there is also

[1] See the chapter by Georges Dionne in this book on empirical evidence about insurance fraud.

[2] Note that a claimant is not fraudulent if he relies in good faith on an erroneous valuation of an apparently competent third party—see Clarke (1997)—. However, insurance may affect fraud in markets for credence goods, i.e., markets where producers may provide unnecessary services to comsumers who are never sure about the extent of the services they actually need. See Darby and Karni (1973) on the definition of credence goods

fraud when a policyholder does not disclose relevant information when he takes out his policy or when he deliberately creates further damages to inflate the size of claim. Lastly, insurance fraud may result from autonomous decision-making of opportunist individuals, but often it goes through collusion with a third party.

Since Becker (1968) and Stigler (1970), the analysis of fraudulent behaviors is part and parcel of economic analysis and there is a growing theoretical literature dealing with insurance fraud. Making progress in this field is all the more important that combating insurance fraud is nowadays a major concern of most insurance companies.

This survey of recent developments in the economic theory of insurance fraud is organized as follows. Sections 2–4 set out the two main approaches to insurance fraud that have been developed in the literature: the costly state verification and the costly state falsification. Both approaches should be considered as complementary. Under the costly state verification hypothesis, the insurer can verify damages but he then incurs a verification (or audit) cost. Under costly state falsification, the policyholder expends some resources for the building-up of his or her claim not to be detected by the insurer. In Section 2, we first describe the general framework used in most parts of our study, namely a model in which a policyholder has private information about the magnitude of his losses and who may file fraudulent claims. We then turn to the analysis of costly state verification procedures under deterministic auditing. In practice, claim handlers are, to some extent, entrusted with claims verification but, more often than not, state verification involves some degree of delegation. Indeed, there are specific agents, such as experts, consulting physicians, investigators or attorneys who are in charge of monitoring claims. Under deterministic auditing, claims are either verified with certainty or not verified at all, according to the size of the claim. The developments in the economic theory of insurance fraud surveyed in Sections 3 and 4 emphasize the fact that policyholders may engage in costly claims falsification activities, possibly by colluding with a third party such as an auto mechanic, a physician or an attorney. Section 3 remains within the costly state verification approach. It is devoted to the analysis of audit cost manipulation: policyholders may expend resources to make the verification of damages more difficult. Section 4 adresses the (*stricto sensu*) costly state falsification approach: at some cost, policyholders are supposed to be able to falsify the actual magnitude of their losses. In other words, they can take acts that misrepresent the actual losses and then the claims' build up cannot

and Dionne (1984) on the effects of insurance on the possibilities of fraud in markets for credence goods.

be detected. Sections 5 to 8 set out extensions of the costly state verification model in various directions. Section 5 focuses on random auditing. Section 6 characterizes the equilibrium of a competitive insurance market where trades are affected by adverse selection because insurers cannot distinguish honest policyholders from potential defrauders. Section 7 focuses on credibility constraints that affect antifraud policies. Section 8 shows that conditioning the decision to audit on fraud signals improves the efficiency of costly stae verification mechanisms and it makes a bridge between auditing and scoring. Section 9 contemplates some indirect effects of insurance contracts on fraud. Sections 10 and 11 focus on collusion, respectively between policyholders and agents in charge of marketing insurance contract in Section 10, and between policyholders and service providers in Section 11. Section 12 concludes. Proofs and references for proofs are gathered in an appendix.

2 Costly state verification: the case of deterministic auditing

Identical insurance buyers own an initial wealth W and they face an uncertain monetary loss x, where x is a random variable with a support $[0, \overline{x}]$ and a cumulative distribution $F(x)$. The no-loss outcome—i.e., the "no-accident" event—may be reached with positive probability. Hence x is distributed according to a mixture of discrete and continuous distributions: x has a mass of probability $f(0)$ at $x = 0$ and there is a continuous probability density function $f(x) = F'(x)$ over $(0, \overline{x}]$. In other words $f(x)/[1 - f(0)]$ is the density of damages conditional on a loss occurring.

The insurance policy specifies the (non negative) payment $t(x)$ from the insurer to the policyholder if the loss is x and the premium P paid by the policyholder. The realization of x is known only to the policyholder unless there is verification, which costs c to the insurer.

For the time being, we assume that the insurer has no information at all about the loss suffered by the policyholder unless he verifies the claim through an audit, in which case he observes the loss perfectly.[3] We will later on consider alternative assumptions, namely the case where the insurer has partial information about the loss suffered (he can costlessly observe whether an accident has occurred but not the magnitude of the loss) and the case where the claim is a fasified image of true damages.

The policyholder's final wealth is $W_f = W - P - x + t(x)$. Policyholders

[3] On imperfect auditing, in contexts which are different from insurance fraud, see Baron and Besanko (1984) and Puelz and Snow (1997).

are risk-averse. They maximize the expected utility of final wealth $EU(W_f)$, where $U(.)$ is a twice differentiable von Neumann-Morgenstern utility function, with $U' > 0$, $U'' < 0$.

A *deterministic auditing policy* specifies whether a claim is verified or not depending on the magnitude of damages. More precisely, following Townsend (1979), we define a deterministic audit policy as a verification set $M \subset [0, \overline{x}]$, with complement M^c, that specifies when there is to be verification. A policyholder who experiences a loss x may choose to file a claim $\widehat{x}$. If $\widehat{x} \in M$, the claim is audited, the loss x is observed and the payment is $t(x)$. If $\widehat{x} \in M^c$, the claim is not audited and the payment to the policyholder is $t(\widehat{x})$.

A contract $\delta = \{t(.), M, P\}$ is said to be *incentive compatible* if the policyholder truthfully reveals the actual loss, i.e., if $\widehat{x} = x$ is an optimal strategy for the policyholder. Lemma 1 establishes that any contract is weakly dominated[4] by an incentive compatible contract, in which the payment is constant in the no-verification set M^c and always larger in the verification set than in the no-verification set.

Lemma 1 *Any contract $\delta = \{t(.), M, P\}$ is weakly dominated by an incentive compatible contract $\widetilde{\delta} = \{\widetilde{t}(.), \widetilde{M}, \widetilde{P}\}$ such that :*

$$\begin{aligned} \widetilde{t}(x) &= t_0 \quad \textit{if} \quad x \in \widetilde{M}^c, \\ \widetilde{t}(x) &> t_0 \quad \textit{if } x \in \widetilde{M}, \end{aligned}$$

where t_0 is some constant.

The characterization of the incentive compatible contracts described in Lemma 1 is quite intuitive. In the first place, truthful revelation of the actual loss is obtained by paying a constant indemnity in the no-verification set, for otherwise the policyholder would always report the loss corresponding to the highest payment in this region. Secondly, if the payment were lower for some level of loss located in the verification set than in the no-verification set, then, for this level of loss, the policyholder would announce falsely that his loss is in the no-verification set.[5]

[4]Dominance is in a Pareto-sense with respect to the expected utility of the policyholder and to the expected profit of the insurer.

[5]If both payments were equal, then it would be welfare improving not to audit the corresponding level of loss in the verification region and simultaneously to decrease the premium. Note that Lemma 1 could be presented as a consequence of the Revelation Principle (see Myerson, 1979).

Lemma 1 implies that we may restrict our characterization of optimal contracts to such incentive compatible contracts. This is proved by defining $\widetilde{t}(x)$ as the highest indemnity payment that the policyholder can obtain when his loss is x, by choosing $\widetilde{M}$ as the subset of $[0,\overline{x}]$ where the indemnity is larger than the minimum and by letting $\widetilde{P}=P$. This is illustrated in Figure 1, with $M=(x^{*},\overline{x}]$, $\widetilde{M}=(x^{**},\overline{x}]$, $\widetilde{t}(x)=t_0$ if $x\leq x^{**}$ and $\widetilde{t}(x)=t(x)$ if $x>x^{**}$. Under δ, for any optimal reporting strategy the policyholder receives t_0 when $x\leq x^{**}$ and he receives $t(x)$ when $x>x^{**}$, which corresponds to the same payment as under $\widetilde{\delta}$. Furthermore, under δ, any optimal strategy $\widehat{x}(x)$ is such that $\widehat{x}(x)\in M$ if $x>x^{**}$, which implies that verification is at least as frequent under δ (for any optimal reporting strategy) as when the policyholder tells the truth under $\widetilde{\delta}$. Thus, δ and $\widetilde{\delta}$ lead to identical indemnity payments whatever the true level of the loss and expected audit costs are lower when there is truthtelling under $\widetilde{\delta}$ than under δ.

From now on, we restrict ourselves to such incentive compatible contracts. The optimal contract maximizes the policyholder's expected utility

$$EU=\int_{M}U(W-P-x+t(x))dF(x)+\int_{M^{c}}U(W-P-x+t_0)dF(x), \quad (1)$$

with respect to $P,t_0,t(.):M\longrightarrow R_{+}$ and $M\subset[0,\overline{x}]$, subject to a constraint that requires the expected profit of the insurer $E\Pi$ to meet some minimum preassigned level normalized at zero

$$E\Pi=P-\int_{M}[t(x)+c]dF(x)+\int_{M^{c}}t_0dF(x)\geq 0, \quad (2)$$

and to the incentive compatibility constraint

$$t(x)>t_0 \text{ for all } x \text{ in } M. \quad (3)$$

Figure 1

Lemma 2 *For any optimal contract, we have*

$$t(x)=x-k>t_0 \textit{ for all } x \textit{ in } M,$$

and

$$M=(m,\overline{x}] \textit{ with } m\in[0,\overline{x}].$$

Lemma 2 shows that it is optimal to verify the claims that exceed a threshold m and also to provide full insurance of marginal losses when $x>m$. The

intuition of these results are as follows. The optimal policy shares the risk between the insured and the insurer without inducing the policyholder to misrepresent his loss level. As shown in Lemma 1, this incentive compatibility constraint implies that optimally the indemnity schedule should be minimal and flat outside the verification set, which means that no insurance of marginal losses is provided in this region. On the contrary, nothing prevents the insurer to provide a larger variable coverage when the loss level belongs to the verification set. Given the concavity of the policyholder's utility function, it is optimal to offer the flat minimal coverage when losses are low and to provide a larger coverage when losses are high. This leads us to define the threshold m that separates the verification set and its complement. Furthermore, conditionally on the claim being verified, i.e., when $x > m$, sharing the risk optimally implies that full coverage of marginal losses should be provided.

Hence, the optimal contract maximizes

$$EU = \int_0^m U(W - x - P + t_0)dF(x) + [1 - F(m)]U(W - P - k),$$

with respect to $P, m \geq 0, t_0 \geq 0$ and $k \geq t_0 - m$ subject to

$$E\Pi = P - t_0 F(m) - \int_{m_+}^{\overline{x}} (c + x - k)dF(x) \geq 0.$$

At this stage it is useful to observe that EU and $E\Pi$ are unchanged if there is a variation in the coverage, constant among states, compensated by an equivalent variation in the premium, i.e., $dEU = dE\Pi = 0$ if $dt_0 = dk = dP$, with m unchanged. Hence, the optimal coverage schedule is defined up to an additive constant. Without loss of generality, we may assume that no insurance payment is made outside the verification set, i.e., $t_0 = 0$. We should then have $t(x) = x - k > 0$ if $x > m$, or equivalently $m - k \geq 0$. In such a case, the policyholder files a claim only if the loss level exceeds the threshold m. This threshold may be viewed as a deductible.

Note that the optimal coverage is no more indeterminate if we assume, more realistically, that the cost c is the sum of the audit cost and of an administrative cost which is incurred whenever a claim is filed, be it verified or not. In such a case, choosing $t_0 = 0$ in the no-verification set is the only optimal solution since it saves the administration cost—see Picard (2000).

The optimal contract is derived by maximizing

$$EU = \int_0^m U(W - x - P)dF(x) + [1 - F(m)]U(W - P - k), \qquad (4)$$

with respect to $m \geq 0, k$ and P, subject to

$$E\Pi = P - \int_{m_+}^{\overline{x}} (c + x + k) dF(x) \geq 0, \tag{5}$$

$$m - k \geq 0. \tag{6}$$

Proposition 1 *Under deterministic auditing, an optimal insurance contract* $\delta = \{t(.), M, P\}$ *satisfies the following conditions:*

$$\begin{aligned} M &= (m, \overline{x}] \text{ with } m > 0, \\ t(x) &= 0 \text{ if } x \leq m, \\ t(x) &= x - k \text{ if } x > m, \end{aligned}$$

with $0 < k < m$.

The optimal contract characterized in Proposition 1—established by Gollier (1987)—is depicted in Figure 2. First, it states that it is optimal to choose a positive threshold m. The intuition is as follows. When $m = 0$, all positive claims are verified and it is optimal to offer full coverage, i.e., $t(x) = x$ for all $x > 0$. Starting from such a full insurance contract an increase $dm > 0$ entails no first-order risk-sharing effect. However, this increase in the threshold cuts down the expected audit cost, which is beneficial to the policyholder. In other words, in the neighbourhood of $m = 0$ the trade-off between cost minimization and risk-sharing always tips in favor of the first objective.

Secondly, we have $0 < k < m$ which means that partial coverage is provided when $x > m$. Intuitively, the coverage schedule is chosen so as to equalize the marginal utility of final wealth in each state of the verification set with the expected marginal utility of final wealth, because any increase in the insurance payment has to be compensated by an increase in the premium paid whatever the level of the loss. We know that no claim is filed when $x < m$, which implies that the expected marginal utility of final wealth is larger than the marginal utility in the no-loss state. Concavity of the policyholder's utility function then implies that a partial coverage is optimal when the threshold is crossed.

Figure 2

Thus far we have assumed that the insurer has no information at all about the loss incurred by the policyholder. In particular, the insurer could not observe whether a loss occurred ($x > 0$) or not ($x = 0$). Following

Bond and Crocker (1997), we may alternately assume that the fact that the policyholder has suffered some loss is publicly observable. The size of the loss remains private information to the policyholder: verifying the magnitude of the loss costs c to the insurer.

This apparently innocuous change in the information structure strongly modifies the shape of the optimal coverage schedule. The insurer now pays a specific transfer $t = t_1$ when $x = 0$, which occurs with probability $f(0)$. Lemmas 1 and 2 are unchanged and we now have

$$EU = f(0)U(W - P + t_1) + \int_{0_+}^{m} U(W - x - P + t_0)dF(x) + [1 - F(m)]U(W - P - k),$$

$$E\Pi = P - t_1 f(0) - t_0[F(m) - f(0)] - \int_{m_+}^{\overline{x}} (c + x - k)dF(x).$$

The optimal contract maximizes EU with respect to $P, m \geq 0, t_0 \geq 0, t_1 \geq 0$ and $k \geq t_0 - m$ subject to $E\Pi \geq 0$. We may choose $t_1 = 0$, since P, t_0, t_1 and k are determined up to an additive constant: no insurance payment is made if no loss occurs.

Proposition 2 *Under deterministic auditing, when the fact that the policyholder has suffered some loss is publicly observable, an optimal insurance contract $\delta = \{t(.), M, P\}$ satisfies the following conditions:*

$$\begin{aligned} &M = (m, \overline{x}] \text{ with } m > 0, \\ &t(0) = 0, \\ &t(x) = t_0 \text{ if } 0 < x \leq m, \\ &t(x) = x \text{ if } x > m, \end{aligned}$$

with $0 < t_0 < m$.

Proposition 2 is established by Bond and Crocker (1997). It is depicted in Figure 3. When an accident occurs but the claim is not verified (i.e., $0 < x \leq m$), the incentive compability requires the insurance payment to be constant: we then have $t(x) = t_0$. The payment should be larger than t_0 when the claim is verified (i.e., when $x > m$). Optimal risk sharing implies that the policyholder's expected marginal utility (conditional on the information of the insurer) should be equal to the marginal utility in the no-accident state. This implies first that, in the no-verification region, an optimal insurance contract entails overpayment of small claims (when $0 < x \leq t_0$) and underpayment of large claims (when $t_0 < x \leq m$). Secondly, there is full insurance in the verification region (i.e., when $x > m$).

Figure 3

Neither Figure 2 nor Figure 3 looks like the coverage schedules that are most frequently offered by insurer for two reasons: first because of the upward discontinuity at $x = m$ and secondly because of overpayment of smaller claims in the case of Figure 3. In fact, such contracts would incite the policyholder to inflate the size of his claim by intentionally increasing the damage. Consider for example the contract described in Proposition 1 and illustrated by Figure 2. A policyholder who suffers a loss x less than m but greater than m' would profit by increasing the damage up to $x = m$, insofar as the insurer is not able to distinguish the initial damage and the extra damage.[6] In such a case, the contract defined in Proposition 1 is dominated by a contract with a straight deductible, i.e., $t(x) = Sup\{0, x-m'\}$ with $M = (m', \overline{x}]$. As shown by Hubermann *et al.* (1983) and Picard (2000), in different settings, a straight deductible is indeed optimal under such circumstances. We thus have:

Proposition 3 *Under deterministic auditing, when the policyholders can inflate their claims by intentionally increasing the damage, the optimal insurance contract* $\delta = \{t(.), M, P\}$ *is a straight deductible*

$$t(x) = Sup\{0, x - m\},$$

with $m > 0$ *and* $M = (m, \overline{x}]$.

Proposition 3 explains why insurance policies with straight deductibles are so frequently offered by insurers, in addition to the wellknown interpretations in terms of transaction costs (Arrow, 1971) or moral hazard (Holmström, 1979).

3 Costly state verification: deterministic auditing with manipulation of audit costs

In the previous section, the policyholder was described as a purely passive agent. His only choices were whether he files a claim or not and, should

[6] In fact, the policyholder would never increase the damage if and only if $t(x) - x$ were non-increasing over $[0, \overline{x}]$. Given that $t(x)$ is non-decreasing (see Lemma 2), this no-manipulability condition implies that $t(x)$ should be continuous. Note that extra damages made either deliberately by the policyholder (arson is a good example) or thanks to a middleman, such as a car repairer or a health case provider. In such cases, gathering verifiable information about intentional overpayment may be too time consuming to the insurer. See Bourgeon and Picard (1999) on corporate fire insurance when there is a risk of arson.

the occasion arise, what is the size of the claim? As a matter of fact, in many cases, the policyholder involved in an insurance fraud case plays a much more active part. In particular, he may try to falsify the damages in the hope of receiving a larger insurance payment. Usually, falsification goes through collusion with agents, such as healthcare providers, car repairers or attorneys, who are in position to make it more difficult or even impossible to prove that the claim has been built up or deliberately created.[7] Even if fraudulent claiming may be detered at equilibrium, the very possibility for policyholders to falsify claims should be taken into account in the analysis of optimal insurance contracts.

Two main approaches to claims falsification have been developed in the literature. Firstly, Bond and Crocker (1997) and Picard (2000) assume that the policyholder may manipulate audit costs, which means that they expend resources to make the verification of claims more costly or more time consuming to the auditor. In this approach, detering the policyholder from manipulating audit cost is feasible and, sometimes, optimal. What is most important is the fact that the coverage schedule affects the incentives of policyholders to manipulate audit costs, which gives a specific moral hazard dimension to the problem of designing an optimal insurance contract. In another approach, developed by Crocker and Morgan (1997), it is assumed that policyholders may expend resources to falsify the actual magnitude of their losses in an environment where verification of claims is not possible. Here also the coverage schedule affects the incentives to claims falsification, but the cost of generating insurance claims through falsification differs among policyholders according to their true level of loss. These differential costs make it possible to implement loss-contingent insurance payments with some degree of claims falsification at equilibrium.

In this section and the following, we review both approaches in turn. For the sake of expositional clarity, we refer to them as costly state verification with manipulation of audit cost and costly state falsification, although in both cases the policyholder falsifies his claim, i.e., he prevents the insurer observing the true level of damages. In the first approach, the policyholder deters the auditor from performing an informative audit while in the second

[7] On collusion between physicians and workers, see the analysis of workers' compensations by Dionne and St-Michel (1991) and Dionne *et al.*(1995). See Derrig *et al.* (1994) on empirical evidence about the effect of the presence of an attorney on the probability of reaching the monetary threshold that restricts the eligibility to file a tort claim in the Massachusetts no-fault automobile insurance system. In the Tort system, Cummins and Tennyson (1992) describe the costs to motorists experiencing minor accidents of colluding with lawyers and physicians as the price of a lottery ticket. The lottery winnings are the motorist's share of a general damage award.

one he provides a distorted image of his damages.

The audit cost manipulation hypothesis has been put forward by Bond and Crocker (1997) in the framework of a model with deterministic auditing. They assume that policyholders may take actions (refered to as *evasion costs*) that affect the audit cost. Specifically, Bond and Crocker assume that, after observing their loss x, a policyholder may incur expenditures $e \in \{e_0, e_1\}$, with $e_1 > e_0$, which randomly affects the audit cost. If $e = e_i$, then the audit cost is $c = c^H$ with probability p_i and $c = c^L$ with probability $1 - p_i$, with $i \in \{0, 1\}, c^H > c^L$ and $p_1 > p_0$. In other words, a large level of manipulation expenditures makes it more likely that the audit cost will be large. Without loss of generality, assume $e_0 = 0$. Let us also simplify by assuming $c^L = 0$. These expenditures are in terms of utility so that the policyholder's utility function is now $U(W_f) - e$.

Bond and Crocker assume that the actual audit cost is verifiable, so that the insurance contract may be conditionned on c. Under deterministic auditing, an insurance contract δ is then defined by a premium P, a state-contingent coverage schedule $t^i(x)$ and a state-contingent verification set $M^i = (m^i, \overline{x}]$, where $i = H$ if $c = c^H$ and $i = L$ if $c = c^L$. Bond and Crocker also assume that the insurer can observe whether an accident has occured, but not the size of the actual damages and (without loss of generality), they assume that no insurance payment is made if $x = 0$.

An optimal *no-manipulation* insurance contract maximizes the expected utility of the policyholder subject to:

- The insurer's participation constraint
- Incentive compatibility constraints that may be written as

$$t^i(x) = \begin{cases} t_0^i & \text{if} \quad x \in (0, m^i] \\ > t_0^i & \text{if} \quad x \in (m^i, \overline{x}] \end{cases}$$

  for $i = H$ or L.

- The constraint that the policyholder does not engage in audit cost manipulation whatever his loss, i.e.,

$$\begin{aligned} &p_1 U(W - x - P + t^H(x)) + (1 - p_1) U(W - x - P + t^L(x)) - e_1 \\ &\quad \leq p_0 U(W - x - P + t^H(x)) + (1 - p_0) U(W - x - P + t^L(x)) \end{aligned}$$

  for all x in $(0, \overline{x}]$.

Bond and Crocker (1997) show the following proposition.

Proposition 4 *The optimal no-manipulation insurance contract* $\delta = \{t^H(.), t^L(.), m^H, m^L, P\}$ *has the following properties:*

(i) $m^H < \overline{x}$ *and* $m^L = 0$

(ii) $t^H(x) = x$ *for* $x > m^H$ *and* $t^H(x) = t_0^H$ *for* $0 < x \leq m^H$

(iii) $t^L(x) = x$ *for* $\widetilde{x} \leq x \leq \overline{x}$ *and* $t^L(x) = S(x)$ *for* $0 < x < \widetilde{x}$ *where* $S(x)$ *is given by*

$$(p_1 - p_0)[U(W - x - P + t_0^H) - U(W - x - P + S(x)] - e_1 = 0.$$

The optimal no-manipulation contract is depicted in Figure 4. If there were no possibility of audit cost manipulation, then the optimal insurance contract would involve $m^L = 0$ and $t^L(x) = x$ for all x (since $c^L = 0$) and $m^H > 0, t^H(x) = x$ if $x > m$ and $0 < t_0^H < m_H$ (see Proposition 2). This suggests that manipulating audit cost (i.e., choosing $e = e_1$) may be a profitable strategy for low values of x. Proposition 4 shows that overcompensating easily verified losses is an appropriate strategy to mitigate the policyholder's incentive to engage in audit cost manipulation. This overcompensation is defined by the $S(x)$ function. $S(x)$ denotes the minimum payoff in the c^L state that makes the policyholder indifferent between manipulating or not and $\widetilde{x}$ is the threshold under which the policyholder chooses to evade if he is offered the full insurance contract in the c^L state.

Since overcompensating is costly to the insurer, it may be optimal to allow for some degree of manipulation at equilibrium. Bond and Crocker provide a characterization of this optimal contract with audit cost manipulation at equilibrium. In particular, they show that there is still a subintervall $[s_2, s_1]$ in $(0, m^H)$ where the insurer overcompensates the loss in the c^L state, with $t^L(x) = S(x) > x$ when $s_2 \leq x < s_1$. Finally they show that, when U exhibits constant absolute risk aversion, then the optimal contract in the presence of audit cost manipulation results in lower payoffs and less monitoring in the c^H state than would an optimal contract in an environment where claims manipulation was not possible.[8]

Figure 4

The analysis of Bond and Crocker (1997) is interesting firstly because it is a first step toward a better understanding of the active part that policyholders may take in insurance fraud. Furthermore, it provides a rationale for the fact

[8]The CARA assumption eliminates wealth effects from incentives constraints.

that insurers may be willing to settle small claims generously and without question when the loss is easily monitored to forestall a claim that may be larger and more difficult to verify. From a normative point of view, the Bond-Crocker analysis suggests that the appropriate way to mitigate build-up is not to increase the amount of monitoring but to design coverage schedules in such a way that policyholders have less incentive to engage in fraudulent claiming.

Two other aspects of the Bond-Crocker model have to be emphasized. First, the optimal coverage schedule is such that small claims are overcompensated whatever the audit cost, which may incite the policyholder to intentionally bring about damages. This issue has already been addressed in Section 3 and we will not hark back to it any further. Secondly, Bond and Crocker assume that the actual audit cost is verifiable so that the insurance coverage may be conditioned on it. This is a very strong assumption. In most cases, claims verification is performed by an agent (an expert, a consulting physician, an attorney, an investigator...) who may have private information about the cost entailed by a specific claim. Picard (2000) focuses attention on the agency relationship that links the insurer and the auditor when policyholders may manipulate audit costs and the insurer does not observe the cost incurred by the auditor. His analysis may be summarized as follows.

The auditor sends a report $\widetilde{x} \in [0, \overline{x}]$ which is an evaluation of the magnitude of the loss. Let $\widetilde{x} = \emptyset$ when no audit is performed. Observing the magnitude of the loss costs c_a to the auditor. The policyholder may incur a manipulation cost e and, in such a case, the cost of elicitating *verifiable* information about the size of the damages become $c_a + be$, where the parameter $b > 0$ characterizes the manipulation technology. Furthermore, verifiable information is necessary to prove that the claim has been build up (i.e., to prove that $x < \widehat{x}$). The insurer does not observe the audit cost. He offers an incentive contract to his auditor to motivate him to gather verifiable information about fraudulent claims. Let t and r be respectively the insurance payment and the auditor's fees. Contracts $T(.)$ and $R(.)$ specify t and r as functions of the auditor's report.[9] We have $t = T(\widetilde{x})$ and $r = R(\widetilde{x})$ where $T(.) : [0, \overline{x}] \cup \emptyset \to R_+$ and $T(.) : [0, \overline{x}] \cup \emptyset \to R$.

The auditor-policyholder relationship is described as a three stage audit game. At stage 0, a loss x, randomly drawn in $[0, \overline{x}]$, is privately observed by the policyholder.[10] At stage 1, the policyholder reports a claim $\widehat{x} \in [0, \overline{x}]$

[9] The payment $R(.)$ is net of standard audit cost c_a.

[10] Contrary to the Bond-Crocker (1997) model, it is assumed that the insurer cannot observe whether an accident has occured, i.e., he cannot distinguish the event $\{x = 0\}$ from $\{x > 0\}$. Furthermore, the manipulation cost e is in monetary terms and not utility terms as in Bond-Crocker (1997).

and he incurs the manipulation cost $e \geq 0$. At stage 2, the claim is audited whenever $\widehat{x} \in M = (m, \overline{x}]$. When $\widehat{x} \in M$, the auditor observes x and he reports $\widetilde{x} \in \{x, \widehat{x}\}$ to the insurer. If $\widetilde{x} = x \neq \widehat{x}$, the auditor incurs the cost $c_a + be$ so that his report incorporates verifiable information. If $\widetilde{x} = \widehat{x}$, the auditor's cost is only c_a. The payments to the policyholder and to the auditor are respectively $T(\widetilde{x})$ and $R(\widetilde{x})$.

In this setting, an allocation is described by $\delta = \{t(.), M, P\}$, with $M = (m, \overline{x}]$ and by $\omega(.) : [0, \overline{x}] \to R$, where $\omega(x)$ is the auditor's equilibrium payoff (net of audit cost) when the loss is equal to x.

Contracts $\{T(.), R(.)\}$ are said to implement the allocation $\{\delta, \omega(.)\}$ if at a perfect equilibrium of the audit game, there is no audit cost manipulation (i.e., $e = 0$ for all x), the claim is verified if and only if $x \in M$ and the net payoffs—defined by $T(.)$ and $R(.)$—are equal to $t(x), \omega(x)$ when the loss is equal to x.[11]

In such a setting, the equilibrium audit cost is $\omega(x) + c_a$ if $x \in M$ and $\omega(x)$ if $x \in M^c$. Furthermore, the auditor's participation constraint may be written as

$$\int_0^{\overline{x}} V(\omega(x)) dF(x) \geq \overline{v}, \tag{7}$$

where $V(.)$ is the auditor's von Neumann-Morgenstern utility function, with $V' > 0, V'' \leq 0$ and $\overline{v}$ is an exogenous reservation utility level.

The optimal allocation $\{\delta, \omega(.)\}$ maximizes the policyholder's expected utility, subject to the insurer's and the auditor's participation constraints and to the constraint that there exist contracts $\{T(.), R(.)\}$ that implement $\{\delta, \omega(.)\}$.

Picard (2000) characterizes the optimal allocation in a setting where the policyholder can inflate their claim by intentionally increasing the damages, which implies that $t(x) - x$ should be nonincreasing (see Section 2). His main result is the following:

Proposition 5 *When the auditor is risk averse ($V'' < 0$), the optimal insurance contract is a deductible with coinsurance for high levels of damages:*

$$\begin{aligned} &t(x) = 0 \quad \textit{if} \quad 0 \leq x \leq m, \\ &t(x) = x - m \ \textit{if} \quad m \leq x \leq x_0, \\ &t'(x) \in (0, 1) \quad \textit{if} \quad x_0 \leq x \leq \overline{x}, \end{aligned}$$

with $0 \leq m < x_0 \leq \overline{x}$ *and* $M = (m, \overline{x}]$.

[11] Picard (2000) shows that allowing for audit cost manipulation (i.e., $e > 0$) at equilibrium is a weakly dominated strategy for the insurer.

Furthermore, the auditor's fees (expressed as function of the size of the claim) are

$$r = r_1 - bt(x) \quad if x > m,$$
$$r = r_0 \quad if \quad x \leq m,$$

where r_0 *and* r_1 *are constant.*

Picard (2000) also gives sufficient conditions for $m > 0$ and $x_0 < \overline{x}$. The contracts characterized in Proposition 5 are depicted in Figure 5. We have $t(x) = 0$ when x is in the no-verification set $[0, m]$. Hence, the threshold m may be interpreted as a deductible under which no claim is filed. In the verification set, there is coinsurance of large losses (i.e., the slope of the coverage schedule is less than one when $x > x_0$). Furthermore, the insurer should pay contingent fees to his auditor: the auditor's fees are (linearly) decreasing in the insurance indemnity payment.

Figure 5

The intuition for these results are as follows. Let $x \in M$. A deviation from truthful revelation of loss without audit cost manipulation (i.e., $\widehat{x} = x, e = 0$) to $\widehat{x} = x' > x, e > 0$ is profitable to the policyholder if $T(x') - e > T(x)$ provided the claim is accepted by the auditor, which implies $R(x') \geq R(x) - be$. Both conditions are incompatible (for all e) if

$$R(x') + bT(x') \leq R(x) + bT(x)$$

For all $x \in M$, we have $t(x) = T(x), \omega(x) = R(x)$. This means that $\omega(x) + bt(x)$ should be nonincreasing for manipulation of audit cost to be detered. In other words, a 1 \$ increase in the indemnity payment should lead at least to a b \$ decrease in the auditor's fees. Because the auditor is risk averse, it would be suboptimal to have $\omega'(x) < -bt'(x)$, which gives the result on contingent fees. Because of condition $\omega'(x) = -bt'(x)$, a greater scope of variation in insurance payments entails a greater variability in the auditor's fees and thus a larger risk premium paid to the auditor for his participation constraint to be satisfied. Some degree of coinsurance for large losses then allows the insurer to decrease the auditor's expected fees which is ultimately beneficial to the policyholder. This argument does not hold if the auditor is risk-neutral and, in that case, a straight deductible is optimal. Inversely, a ceiling on coverage is optimal when the auditor is infinitely risk-averse or when he is affected by a limited liability constraint.

4 Costly state falsification

Let us come now to the analysis of state falsification first examined by Lacker and Weinberg (1989)[12] and applied to an insurance setting by Crocker and Morgan (1997).[13], The policyholders are in position to misrepresent their actual losses by engaging in costly falsification activities. The outcome of these activities is a claim denoted by $y \in R_+$. The insurer only observes y: contrary to the costly state verification setting, verifying the actual magnitude of damages is supposed to be prohibively costly. Hence, an insurance contract only specifies a coverage schedule $t = T(y)$. Claims falsification is costly to the policyholder, particularly because it may require colluding with a provider (an automechanics, a physician...) or using the services of an attorney. Let $C(x, y)$ be the falsification cost. The policyholder's final wealth becomes

$$W_f = W - x - P + T(y) - C(y, x).$$

Let $y(x)$ be the (potentially falsified) claim of a policyholder who suffers an actual loss x. Given a falsification strategy $y(.) : [0, \overline{x}] \to R_+$, the policyholder's final wealth may be written as a function of his loss:

$$W_f(x) \equiv W - x - P + T(y(x)) - C(y(x), x) \tag{8}$$

An optimal insurance contract maximizes $EU(W_f(x))$ with respect to $T(.)$ and P subject to

$$P \geq \int_0^{\overline{x}} T(y(x))dF(x), \tag{9}$$

$$y(x) \in \text{Arg Max}_{y'} T(y') - C(y', x) \quad \text{for all } x \in [0, \overline{x}]. \tag{10}$$

(9) is the insurer's participation constraint and (10) specifies that $y(x)$ is an optimal falsification strategy of a type-x policyholder.

Since the payments $\{P, T(.)\}$ are defined up to an additive constant, we may assume $T(0) = 0$ without loss of generality. For the time being, let us restrict attention to linear coverage schedule, i.e., $T(y) = \alpha y + \beta$. Our normalization rule gives $\beta = 0$. Assume also that the falsification costs borne by the policyholder depend upon the absolute amount of misrepresentation $(y - x)$ and, for the sake of simplicity, assume $C = \gamma(y - x)^2/2$, where γ is an exogenous cost parameter. (10) then gives

$$y(x) \equiv x + \frac{\alpha}{\gamma}. \tag{11}$$

[12] See also Maggi and Rodriguez-Clare (1995).

[13] Hau (2008) analyzes costly state verification and costly state falsification in a unified model. See Crocker and Tennyson (1999),(2002) and Dionne and Gagné (2001) on econometric testing of the theoretical predictions of models involving costly state falsification.

Hence the amount of falsification $y(x) - x$ is increasing in the slope of the coverage schedule and decreasing in the falsification cost parameter. The optimal coverage schedule will tradeoff two conflicting objectives: providing more insurance to the policyholder, which requires increasing α, and mitigating the incentives to claim falsification by lowering α.

The insurer's participation constraint (9) is binding at the optimum, which gives

$$P = \int_0^{\overline{x}} \left(\alpha x + \frac{\alpha^2}{\gamma} \right) dF(x) = \alpha Ex + \frac{\alpha^2}{\gamma}.$$

(8) then gives

$$W_f(x) = W - (1-\alpha)x - \alpha Ex + \frac{\alpha^2}{2\gamma}$$

Maximizing $EU(W_f(x))$ with respect to α leads to the following first-order condition

$$\frac{\partial EU}{\partial \alpha} = E\left\{ \left(x - Ex - \frac{\alpha}{\gamma} \right) U'(W_f(x)) \right\} = 0, \tag{12}$$

and thus

$$\frac{\partial EU}{\partial \alpha} \mid_{\alpha=1} = -\frac{1}{\gamma} U' \left(W - Ex - \frac{1}{2\gamma} \right) < 0, \tag{13}$$

$$\frac{\partial EU}{\partial \alpha} \mid_{\alpha=0} = E\{(x - Ex)U'(W - x)\} > 0. \tag{14}$$

We also have

$$\frac{\partial^2 EU}{\partial \alpha^2} = -\frac{1}{\gamma} EU'(W_f(x)) + E\left\{ \left(x - Ex - \frac{\alpha}{\gamma} \right)^2 U''(W_f(x)) \right\} < 0, \tag{15}$$

which implies that $0 < \alpha < 1$ at the optimum. Hence, under costly state falsification, the optimal linear coverage schedule entails some degree of coinsurance and (11) shows that there exists a certain amount of claims falsification at equilibrium. This characterization results from the trade-off between the above mentioned confliting objectives: providing insurance to the policyholder and detering him from engaging in costly claim falsification activities.

This trade-off is particularly obvious when $U(.)$ is quadratic. In that case, we may write

$$EU(W_f) = EW_f - \eta Var(W_f) \quad \text{with } \eta > 0 \tag{16}$$

and straightforward calculations give

$$\alpha = \frac{2\eta\gamma\sigma^2}{1 + 2\eta\gamma\sigma^2} \tag{17}$$

at the optimum, where $\sigma^2 \equiv Var(x)$.

Hence, the coinsurance coefficient α is an increasing function of the cost parameter γ, of the risk aversion index η and of the variance of the loss. We have

$$T(y(x)) = \alpha x + \frac{\alpha^2}{\gamma},$$

which give $T(y(x)) > x$ if $x < x_0$ and $T(y(x)) < x$ if $x > x_0$ with $x_0 = \alpha^2/\gamma(1-\alpha)$. Hence in this case, the optimal indemnification rule overcompensates small losses and it overpays larger ones. This is depicted in Figure 6.

Assume now that the insurer observes whether a loss occured or not, as in the paper by Crocker and Morgan (1997). Then an insurance contract is defined by a premium P, an insurance payment t_0 if $x = 0$ and an insurance coverage schedule $T(y)$ to be enforced if $x > 0$. In that case, a natural normalization rule is $t_0 = 0$. We still assume that $T(y)$ is linear: $T(y) = \alpha y + \beta$. For the sake of simplicity, we also assume that $U(.)$ is quadratic.

Figure 6

The insurer's participation constraint and (11) give

$$P = \alpha Ex + [1 - f(0)] \left(\frac{\alpha^2}{\gamma} + \beta \right), \tag{18}$$

which implies

$$W_f = W - \alpha Ex - [1 - f(0)] \left(\frac{\alpha^2}{\gamma} + \beta \right) \quad \text{if} \quad x = 0,$$

$$W_f = W - \alpha Ex - [1 - f(0)] \left(\frac{\alpha^2}{\gamma} + \beta \right) - (1-\alpha)x + \beta + \frac{\alpha^2}{2\gamma} \quad \text{if} \quad x > 0,$$

and we obtain

$$EW_f = W - Ex - \frac{\alpha^2}{2\gamma}[1 - f(0)], \tag{19}$$

and

$$Var(W_f) = f(0)[1-f(0)] \left(\beta + \frac{\alpha^2}{2\gamma} \right)^2 + (1-\alpha)^2\sigma^2 - 2f(0)(1-\alpha) \left(\beta + \frac{\alpha^2}{2\gamma} \right) Ex. \tag{20}$$

Maximizing $EU(W_f)$ defined by (16) with respect to α and β gives the following result

$$\alpha = \frac{2\eta\gamma\widetilde{\sigma}^2}{1 + 2\eta\gamma\widetilde{\sigma}^2}, \tag{21}$$

$$\beta = (1-\alpha)\overline{x} - \frac{\alpha^2}{2\gamma}, \tag{22}$$

where $\widetilde{\sigma}^2 = Var(x \mid x > 0)$ and $\overline{x} = E(x \mid x > 0)$ i.e., $\widetilde{\sigma}^2$ and $\overline{x}$ are respectively the variance and the expected value of the magnitude of damages conditional on a loss occuring.

(21) is similar to (17) and it may be interpreted in the same way. The fact that α is strictly positive (and less than one) means that some degree of insurance is provided but also that there is claims falsification at equilibrium. β may be positive or negative, but the insurance payment $T(y(x))$ is always positive.[14] As in the previous case, small losses are overcompensated and there is undercompensation for more severe losses.

Crocker and Morgan (1997) obtain a similar characterization without restricting themselves to a linear-quadratic model. They characterize the allocations, $\{t(.), y(.), P\}$, with $t(.) : [0, \overline{x}] \to R_+$ and $y(.) : (0, \overline{x}] \to R_+$, that may be implemented by a coverage schedule $T(y)$.[15] For such an allocation, there exists $T(.) : R_+ \to R_+$ such that

$$y(x) \in \text{Arg max}_{y'} \{T(y') - C(y', x)\},$$

and

$$t(x) = T(y(x)) \quad \text{for all } x.$$

The Revelation Principle (Myerson, 1979) applies in such a context, which means that implementable allocations may be obtained as the outcome of a revelation game in which

1. The insurance payment t and the action y are defined as functions of a message $\widetilde{x} \in [0, \overline{x}]$ of the policyholder, i.e., $t = t(\widetilde{x}), y = y(\widetilde{x})$.

2. Truthtelling is an optimal strategy for the policyholder, i.e.,

$$x \in \text{Arg Max}_{\widetilde{x}} \{t(\widetilde{x} - C(y(\widetilde{x}), x)\} \tag{23}$$

[14] When β is negative, the optimal coverage schedule is equivalent to a deductible $m = -\beta/\alpha$ with a coinsurance provision for larger losses, i.e., $T(y(x)) = Sup\{0, \alpha(y - m)\}$.

[15] Crocker and Morgan assume that the insurer can observe whether a loss occurred or not. Hence, there may be falsification only if $x > 0$.

for all x in $(0, \overline{x}]$.

Such an allocation $\{t(.), y(.)\}$ is said to be incentive compatible. The optimal allocation maximizes the policyholder's expected utility $EU(W_f(x))$ with respect to $t(.), y(.)$ and P subject to the insurer's participation constraint and to incentive compatibility constraints. Using a standard technique of incentives theory, Crocker and Morgan characterize the optimal solution of a less-constrained problem in which a first-order truthtelling condition is substituted to (23). They obtain the following result.[16],[17]

Proposition 6 *The optimal solution to the insurance problem under claims falsification satisfies*

$$\begin{aligned} &y(0_+) = 0, y(\overline{x}) = \overline{x} \text{ and } y(x) > x \text{ if } 0 < x < \overline{x}, \\ &t'(0_+) = t'(\overline{x}) = 0 \text{ and } t'(x) > 0 \text{ if } 0 < x < \overline{x}, \\ &t(0_+) > 0 \text{ and } t(\overline{x}) < \overline{x}. \end{aligned}$$

Proposition 6 extends the results already obtained in this section to a more general setting, with a non linear coverage schedule. The optimal solution always entails some degree of falsification except at the top (when $x = \overline{x}$) and at the bottom (when $x \to 0_+$). The insurance payment is increasing in the magnitude of the actual damages and it provides overinsurance (respect. underinsurance) for small (respect. large) losses.

5 Costly state verification: the case of random auditing

We now come back to the costly state verification setting. Under *random auditing*, the insurer verifies the claims with a probability that depends upon the magnitude of damages. The insurance payment may differ depending on whether the claim has been verified or not. A policyholder who suffers a loss x files a claim $\widehat{x}$ that will be audited with probability $p(\widehat{x})$. If there is an audit, the true damages are observed by the insurer and the policyholder

[16] There are some minor differences between the Crocker-Morgan's setting and ours. They are not mentioned for the sake of brevity

[17] The second-order condition for incentive compability requires $y(x)$ to be monotonically increasing. If the solution to the less constrained problem satisfies this monotonicity condition, then the optimal allocation is characterized as in Proposition 6. See Crocker and Morgan (1997) for a numerical example. If this is not the case, then the optimal allocation entails bunching on (at least) an interval $(x', x'') \subset [0, \overline{x}]$, i.e., $y(x) = \widehat{y}, t(x) = \widehat{t}$ for all x in (x', x''). In such a case, the coverage schedule $T(y)$ that sustains the optimal allocation is not differentiable at $y = \widehat{y}$.

receives an insurance payment $t_A(x,\widehat{x})$. If there is no audit, the insurance payment is denoted $t_N(\widehat{x})$.

When a policyholder with damages x files a claim $\widehat{x}$, is expected utility is

$$[1-p(\widehat{x})]U(W-P-x+t_N(\widehat{x}))+p(\widehat{x})U(W-P-x+t_A(x,\widehat{x})).$$

The *Revelation Principle* applies to this setting and we can restrict attention to incentive compatible insurance contracts, that is to contracts where the policyholder is given incentives to report his loss truthfully. Such incentive compatible contracts are such that

$$\begin{aligned} & [1-p(x)]U(W-P-x+t_N(x))+p(x)U(W-P-x+t_A(x,x)) \\ \geq\ & [1-p(\widehat{x})]U(W-P-x+t_N(\widehat{x}))+p(\widehat{x})U(W-P-x+t_A(x,\widehat{x})) \end{aligned} \quad (24)$$

for all $x,\widehat{x}\neq x$.

Let us assume that the net payment from the policyholder to the insurer $P-t_A(x,\widehat{x})$ is bounded by a maximal penalty that can be imposed in case of misrepresentation of damages (i.e., when $x\neq\widehat{x}$). This maximal penalty[18] may depend on the true level of damages x and will be denoted $B(x)$. Hence, we have

$$P-t_A(x,\widehat{x})\leq B(x) \text{ if } x\neq\widehat{x}. \quad (25)$$

For instance, Mookherjee and Png (1989) assume that the wealth of the policyholder is prefectly liquid and that his final wealth can be at most set equal to zero in case of false claim detected by audit. We have $B(x)\equiv W-x$ in that case. Fagart and Picard (1999) assume that the policyholder is affected by a liquidity constraint and that the liquid assets of the policyholder have a given value B. The maximal penalty is then $B(x)=B$ for all x. Another interpretation of (25) is that $B(x)\equiv B$ is an exogenously given parameter that represents the cost (in monetary terms) incurred by a policyholder who is prosecuted after he filed a fraudulent claim detected by audit.[19]

[18] The Revelation Principle does not apply any more if the maximal penalty also depend on the claim $\widehat{x}$. In such a case, there may be false report at equilibrium.

[19] Under this interpretation, it may be more natural to assume that the policyholder should pay the penalty B in addition to the premium P, since the latter is usually paid at the beginning of the time period during which the insurance policy is enforced. In fact, both assumptions are equivalent when the policyholder is affected by a liquidity constraint. Indeed, in such a case, it would be optimal to fix the insurance premium P at the largest possible level (say $P=\overline{P}$) and to compensate adequately the policyholder by providing large insurance payments t_N and t_A unless a fraudulent claim is detected by audit. This strategy provides the highest penalty in case of fraud, without affecting equilibrium net payments t_N-P and t_A-P. If the law of insurance contracts specifies a penalty $\widehat{B}$ to be paid in case of fraudulent claim, we have $P-t_A(x,\widehat{x})\leq\overline{P}+\widehat{B}$ which corresponds to (25) with $B(x)\equiv\overline{P}+\widehat{B}$.

This upper bound on the penalty plays a crucial role in the analysis of optimal insurance contracts under random auditing. Indeed, by increasing the penalty, the insurer could induce truthtelling by the policyholder with a lower probability of auditing, which, since auditing is costly, reduces the cost of the private information. Consequently, if there were no bound on the penalty, first-best optimality could be approximated with very large fines and a very low probability of auditing. Asymetry of information would not be a problem in such a case.

In equilibrium, the policyholder always reports his loss truthfully. Hence, it is optimal to make the penalty as large as possible since this provides maximum incentive to tell the truth without affecting the equilibrium payoffs.[20] We thus have

$$t_A(x, \widehat{x}) = P - B(x) \text{ if } x \neq \widehat{x}$$

Finally, we assume that the policyholder's final wealth W_f should be larger than a lower bound denoted $A(x)$. This bound on the policyholder's final wealth may simply result from a feasibility condition on consumption. In particular, we may have $W_f \geq 0$ which gives $A(x) = 0$ for all x. The lower bound on final wealth may also be logically linked to the upper bound on the penalty: when $B(x)$ corresponds to the value of liquid assets of the policyholder, we have $P - t_N(x) \leq B(x)$ and $P - t_A(x, x) \leq B(x)$ for all x which implies $W_f \geq W - x - B(x) \equiv A(x)$. Mookherjee and Png (1989) assume $B(x) = W - x$, which gives $A(x) = 0$. Fagart and Picard (1999) assume $B(x) = B$, which gives $A(x) = W - x - B$.

Let $t_A(x) \equiv t_A(x, x)$. Under random auditing, a contract will be denoted $\delta = \{t_A(.), t_N(.), p(.), P\}$. An optimal contract maximizes

$$EU = \int_0^{\overline{x}} \{[1-p(x)]U(W-P-x+t_N(x))+p(x)U(W-P-x+t_A(x))\}dF(x) \tag{26}$$

with respect to $P, t_A(.), t_N(.)$ and $p(.)$ subject to the following constraints:

$$E\Pi = P - \int_0^{\overline{x}} \{[1-p(x)]t_N(x) + p(x)[t_A(x)+c]\}dF(x) \geq 0, \tag{27}$$

[20] In a more realistic setting, there would be several reasons for which imposing maximal penalties on defrauders may not be optimal. In particular, audit may be imperfect so that innocent individuals may be falsely accused. Furthermore, a policyholder may overestimate his damages in good faith. Lastly, very large fines may create incentives for policyholders caught cheating to bribe the auditor to overlook their violation.

$$
\begin{aligned}
&[1-p(x)]U(W-P-x+t_N(x))+p(x)U(W-P-x+t_A(x)) \\
&\geq \\
&[1-p(\widehat{x})]U(W-P-x+t_N(\widehat{x}))+p(\widehat{x})U(W-x-B(x)) \\
&\text{for all } x,\widehat{x}\neq x, && (28)\\
&W-P-x+t_N(x)\geq A(x) \text{ for all } x, && (29)\\
&W-P-x+t_A(x)\geq A(x) \text{ for all } x, && (30)\\
&0\leq p(x)\leq 1 \text{ for all } x. && (31)
\end{aligned}
$$

(27) is the insurer's participation constraint. Inequalities (28) are the incentive compatibility constraints that require the policyholder to be willing to report his level of loss truthfully. (29), (30) and (31) are feasibility constraints.[21]

Mookherjee and Png (1989) have established a number of properties of an optimal contract. They are synthetized in Proposition 7 hereafter. In this proposition $\nu(x)$ denotes the expected utility of the policyholder when his loss is x, i.e.,

$$\nu(x)=[1-p(x)]U(W-P-x+t_N(x))+p(x)U(W-P-x+t_A(x)).$$

Proposition 7 *Under random auditing, an optimal insurance contract* $\delta=\{t_A(.),t_N(.),p(.),P\}$*, has the following properties:*

(i) $p(x)<1$ *for all* x *if* $\nu(x)>U(W-x-B(x))$ *for all* x,

(ii) $t_A(x)>t_N(x)$ *for all* x *such that* $p(x)>0$,

(iii) If $p(\widehat{x})>0$ *for some* $\widehat{x}$ *then there exists* x *such that* $\nu(x)=[1-p(\widehat{x})]U(W-x-P+t_N(\widehat{x}))+p(\widehat{x})U(W-x-B(x))$,

(iv) If $\nu(x)>u(W-x-B(x))$ *for all* x *and* $t_N(\widehat{x})=Min\{t_N(x),x\in[0,\overline{x}]\}$, *then* $p(\widehat{x})=0$ *and* $p(x'')>p(x')$ *if* $t_N(x'')>t_N(x')$.

In Proposition 7, the condition "$\nu(x)>U(W-x-B(x))$ for all x" means that nontrivial penalties can be imposed on those detected to have

[21] Deterministic auditing may be considered as a particular case of random auditing where $p(x)=1$ if $x\in M$ and $p(x)=0$ if $x\in M^c$, and Lemma 1 may be obtained as a consequence of the incentive compability conditions (28). If $x,\widehat{x}\in M^c$, (28) gives $t_N(x)\geq t_N(\widehat{x})$. Interverting x and $\widehat{x}$ gives $t_N(\widehat{x})\geq t_N(x)$. We thus have $t_N(x)=t_0$ for all x in M^c. If $x\in M$ and $x\in M^c$, (28) gives $t_A(x)\geq t_N(\widehat{x})=t_0$. If $t_A(x)=t_0$ for $x\in[a,b]\subset M$, then it is possible to choose $p(x)=0$ if $x\in[a,b]$, and to decrease P, the other elements of the optimal contract being unchanged. The policyholder's expected utility would increase, which is a contradiction. Hence $t_A(x)>t_0$ if $x\in M$.

filed a fraudulent claim. Let us call it "condition **C**". Mookherkjee and Png (1989) assume $B(x) = W - x$, which means that the final wealth can be set equal to zero if the policyholder is detected to have lied. In such a case, **C** means that the final wealth is always positive at the optimum and a sufficient condition for **C** to hold is $U'(0_+) = +\infty$. If we assume $B(x) = B$, i.e., the penalty is upward bounded either because of a liquidity constraint or because of statutory provisions, then **C** holds if B is large enough.[22] If **C** does not hold at equilibrium, then the optimal audit policy is deterministic and we are back to the characterization of Section 2. In particular, the $B = 0$ case reverts to deterministic auditing.

From (i) in Proposition 4, all audits must be random if **C** holds. The intuition for this result is that under **C**, the policyholder would always strictly prefer not to lie if his claim were audited with probability one. In such a case, decreasing slightly the audit probability reduces the insurer's expected cost. This permits a decrease in the premium P, and thus an increase in the expected utility of the policyholder, without inducing the latter to lie. (ii) shows that the policyholder who has been verified to have reported his damages truthfully should be rewarded. The intuition is as follows. Assume $t_A(x) < t_N(x)$ for some x. Let $t_A(x)$ –respect. $t_N(x)$– be increased (respect. decreased) slightly so that the expected cost $p(x)t_A(x) + [1 - p(x)]t_N(x)$ is unchanged. This change does not disturb the incentive compatibility constraints and it increases the expected utility which contradicts the optimality of the initial contract. If $t_A(x) = t_N(x)$, the same variation exerts no first-order effect on the expected utility (since we start from a full insurance position) and it allows the insurer to reduce $p(x)$ without disturbing any incentive compatibility constraint. The expected cost decreases, which enables a decreases in the premium P and thus generates an increase in the expected utility. This also contradicts the optimality of the initial contract. (iii) shows that for any level of loss $\widehat{x}$ audited with positive probability, there exists a level of loss x such that the policyholder who suffers the loss x is indifferent between filing a truthful claim and reporting $\widehat{x}$. In other words, when a claim $\widehat{x}$ is audited with positive probability, a decrease in the probability of audit $p(\widehat{x})$ would induce misreporting by the policyholder for (at least) one level of loss x. Indeed, if this were not the case, then one could lower $p(\widehat{x})$ without disturbing any incentive compatibility constraint. This variation allows the insurer to save on audit cost and it enables a decrease in the premium. The policyholder's expected utility increases which contradicts the optimality of the initial contract. Finally, (iv) shows that, under **C**, the claim corresponding to the lowest indemnity payment in the absence of audit should not be

[22] See Fagart and Picard (1999).

audited. All other claims should be audited and the larger the indemnity payment in the absence of audit, the larger the probability of audit. Once again, the intuition is rather straightforward. A policyholder who files a fraudulent claim $\widehat{x}$ may be seen as a gambler who wins the prize $t_N(\widehat{x})$ if he has the luck not to be audited and who will pay $B(x)$ if he gets caught. The larger the prize, the larger the audit probability should be for fraudulent claiming to be detered. Furthermore it is useless to verify the claims corresponding to the lowest prize since it always provides a lower expected utility than truthtelling.

The main difficulty if one wants to further characterize the optimal contract under random auditing is to identify the incentive compatibility constraints that are binding at the optimum and those that are not binding. In particular, it may be that, for some levels of damages, many (and even all) incentive constraints are binding and, for other levels of damages none of them are binding.[23] Fagart and Picard (1999) provide a full characterization of the optimal coverage schedule and of the audit policy when the policyholder has constant absolute risk aversion and the penalty is constant (i.e., $B(x) \equiv B$).

Proposition 8 *Assume $U(.)$ exhibits constant absolute risk aversion and **C** holds at the optimum. Then there exist $m > 0$ and $k \in (0, m)$ such that*

$$\begin{aligned} t_A(x) &= x - k \text{ and } t_N(x) = x - k - \eta(x) \text{ if } x > m \\ t_A(x) &= t_N(x) = 0 \text{ if } x \leq m \end{aligned}$$

with $\eta(x) > 0, \eta'(x) < 0, \eta(m) = m - k, \eta(x) \to 0$ when $x \to \infty$.

Furthermore, we have

$$\begin{aligned} 0 &< p(x) < 1, p'(x) > 0, p''(x) < 0 \text{ when } x > m \\ p(m) &= 0 \\ p(x) &\to \overline{p} \in (0,1) \text{ when } x \to \infty \end{aligned}$$

The optimal contract characterized in Proposition 8 is depicted in Figure 7. No claim is filed, when the magnitude of damages is less than m. When the damages exceed the treshold, then the insurance payment is positive and it is larger when the claim is audited than when it is not—which confirms Proposition 7-(ii)—. However the difference is decreasing when the magnitude of damages is increasing and this difference goes to zero when the

[23] Technically, this rules out the possibility of taking up the differential approach initially developed by Guesnerie and Laffont (1984) and widely used in the literature on incentives contracts under adverse selection.

damages go to infinity (when $\overline{x} = +\infty$). Marginal damages are fully covered in case of audit, i.e., $t'_A(x) = 1$ if $x > m$. In other words, the insurance coverage includes a constant deductible k if the claim is verified. If the claim is not verified, then there is also an additional deductible that disappears when the damages become infinitely large. Furthermore the probability of audit is a concave increasing function of the damages and this probability goes to a limit $\overline{p} < 1$ when x goes to infinity.

Figure 7

To understand the logic of these results, observe that any variation in insurance payment (with a compensating change in the premium) entails two effects. Firstly, it affects the risk sharing between the insurer and the policyholder and, of course, this is the *raison d'être* of any insurance contract. Secondly, it may also modify the audit policy for incentive compatibility constraints not to be disturbed. This second effect is more difficult to analyze because the effects of variations in insurance payment on the incentive to tell the truth are intricate. As above, we may describe the decision making of the policyholder as if he were a gambler. When the true level of damages is x, filing a fraudulent claim $\widehat{x} \neq x$ amounts to choose the lottery "earning $t_N(\widehat{x})$ with probability $1-p(\widehat{x})$ or loosing B with probability $p(\widehat{x})$" in preference to the lottery "earning $t_N(x)$ with probability $1-p(x)$ or earning $t_A(x)$ with probability $p(x)$". If the incentive compability constraint corresponding to x and $\widehat{x}$ is tight, then any increase in $t_N(\widehat{x})$ should be accompanied by an increase in $p(\widehat{x})$ for fraudulent claiming to be detered. However, simultaneously, the increase in $t_N(\widehat{x})$ may also affect the optimal strategy of a policyholder who has actually experienced a loss $\widehat{x}$ and who (for instance) intended to file another fraudulent claim, say $\widehat{x}' \neq \widehat{x}$. This policyholder may come back to truthfulling after the increase in $t_N(\widehat{x})$, even if $t_N(\widehat{x}')$ is slightly increased. This sequence is possible if the preferences of our gambler over lotteries depend upon his wealth, i.e., upon the magnitude of his loss. This suggests that, without simplifying assumptions, analyzing the consequences of a variation in the coverage schedule on the policyholder's strategy may be quite intricate.

The problem is much more simple under constant absolute risk aversion since wealth effects disappear from the incentive constraints when utility is exponential. Fagart and Picard (1999) have considered this case. They show that, when $U(.)$ is CARA, the only incentive constraints that may be binding at the optimum correspond to loss levels $x \in I \subset [0, \overline{x}]$ for which the policyholder receives the smallest indemnity payment. This results from

the fact that, when $U(.)$ is CARA, the loss x disappears from (28). We know from Proposition 7-*(ii)* and *(iv)* that the claim is not audited in that case, which allows us to assume $t_N(x) = t_A(x) = 0$ if $x \in I$ since, as before, the optimal insurance coverage schedule $\{t_N(.), t_A(.), P\}$ is defined up to an additive constant. The best risk-sharing is reached when $I = [0, m]$, with $m > 0$. Under constant absolute risk aversion, the fact that small claims should not be audited can thus be extended to the case of random auditing.

When the loss exceeds m, it is optimal to provide a positive insurance payment. Any increase in $t_N(x)$ should be accompanied by an increase in $p(x)$ for fraudulent claiming to be detered. Let $\Phi(t_N)$ be the probability of audit for which the lottery "earning $t_N(x)$ with probability $1-p(x)$ or loosing B with the probability $p(x)$" and the status quo (i.e., a zero certain gain) are equivalent for the policyholder when his true loss level $\widetilde{x}$ is in I. The probability $\Phi(t_N)$ does not depend on $\widetilde{x}$ when $U(.)$ is CARA and we have $\Phi' > 0$, $\Phi'' < 0$. The optimal audit probability is such that $p(x) = \Phi(t_N(x))$ for all $x > m$.

Let $c\Phi'(t_N(x))dt_N(x)$ be the additional expected audit cost induced by a marginal increase in the insurance payment $dt_N(x)$. Adding this additional expected audit cost to the variation in the insurance payment itself gives the additional expected total cost $[1 + c\Phi'(t_N(x))]dt_N(x)$. When a claim is audited, the additional cost induced by an increase in the insurance payment is just $dt_A(x)$. The difference in additional cost per \$ paid as coverage explains why a larger payment should be promised in case of audit—i.e., $t_A(x) > t_N(x)$—. More precisely, $\Phi'' < 0$ implies that $1 + c\Phi'(t_N(x))$ is decreasing when $t_N(x)$ is increasing. Hence, the difference in the additional expected cost per \$ paid as coverage decreases when $t_N(x)$ increases. This exlains why the additional deductible $t_A(x) - t_N(x) \equiv \eta(x)$ is decreasing and disappears when x is large.[24]

6 Moral standards and adverse selection

Thus far we have assumed that the policyholders are guided only by self-interest and that they didn't feel any morale cost after filing a fraudulent claim. In other words, there was no intrinsic value of honesty to policyholders. In the real world, thank God, dishonesty creates morale problems and a lot of people are detered to file fraudulent claim even if the probability of being caught is small and the fine is moderate. However, more often than

[24] Let $\overline{U}(x) = [1 - p(x)]U(W - P - x + t_N(x)) + p(x)U(W - P - x + t_A(x))$ be the expected utility of a policyholder who has incurred a loss x. Using $p(m) = 0$ shows that $\overline{U}(x)$ is continuous at $x = m$.

not, the insurers are unable to observe the morale cost incurred by their customers which lead to an adverse selection problem.[25] In such a situation, the optimal audit policy as well as the competitive equilibrium in the insurance market (in terms of coverage and premium) may be strongly affected by the distribution of morale costs in the population of policyholders. In particular, the consequences of insurance fraud will be all the more severe that the proportion of purely opportunistic policyholders (i.e., individuals without any morale cost) is large.

We will approach this issue in the following setting, drawn from Picard (1996).[26] Assume that the insurance buyers face the possibility of a loss L with probability $\delta \in (0,1)$. Hence, for the sake of simplicity, the size of the loss is now given. The insurance contract involves a premium P and a level of coverage t. The insurer audits claims with a probability $p \in [0,1]$ at cost c. To simplify further the analysis, we assume that the insurance payment t is the same, whether the claim is audited or not. The reservation utility is $\overline{U} = \delta U(W-L)+(1-\delta)U(W)$. The policyholders may be either opportunist, with probability θ or honest with probability $1-\theta$, with $0 < \theta < 1$. Honest policyholders truthfully report losses to their insurer: they would suffer very large morale cost when cheating. Opportunists may choose to fraudulently report a loss. Let α be the (endogenously determined) probability for an opportunist to file a fraudulent claim when no loss has been incurred. The insurers cannot distinguish honest policyholders from opportunists.

Law exogenously defines the fine, denoted B, that has to be paid by a policyholder who is detected to have lied. Let $\widetilde{p}$ denote the audit probability that makes an opportunist (who has not experienced any loss) indifferent between honesty and fraud. Honesty gives $W_f = W - P$ where W (respect. W_f) still denotes the initial (respect. final) wealth of the policyholder. Fraud gives $W_f = W-P-B$ if the claim is audited and $W_f = W-P+t$ otherwise. Hence $\widetilde{p}$ is given by

$$U(W-P) = \widetilde{p}U(W-P-B) + (1-\widetilde{p})U(W-P+t),$$

which implies

$$\widetilde{p} = \frac{U(W-P+t)-U(W-P)}{U(W-P+t)-U(W-P-B)} \equiv \widetilde{p}(t,P) \in (0,1).$$

Consider a contract (t,P) chosen by a population of individuals that includes a proportion $\sigma \in [0,1]$ of opportunists. Note that σ may conceivably differ

[25] This asymetric information problem may be mitigated in a repeated relationship framework.

[26] See also Boyer (1999) for a similar model.

from θ if various contracts are offered on the market. Given (q, P, σ), the relationship between a policyholder and his insurer is described by the following three stage game:

- At stage 1, *nature* determines whether the policyholder is honest or opportunist, with probabilities $1-\sigma$ and σ respectively. Nature also determines whether the policyholder experiences a loss with probability δ.
- At stage 2, the *policyholder* decides to file a claim or not. Honest customers always tell the truth. When no loss has been incurred, opportunists defraud with probability α.
- At stage 3, when a loss has been reported at stage 2, the insurer audits with probability p.

Opportunists who do not experience any loss choose α to maximize

$$EU = \alpha[pU(W-P-B) + (1-p)U(W-P+t)] + (1-\alpha)U(W-P),$$

which gives

$$\left.\begin{array}{cc} \alpha = 0 & \text{if } p > \widetilde{p}(t,P), \\ \alpha \in [0,1] & \text{if } p = \widetilde{p}(t,P), \\ \alpha = 1 & \text{if } p < \widetilde{p}(t,P). \end{array}\right\} \tag{32}$$

The insurer chooses p to maximize its expected profit $E\Pi$ or equivalently to minimize the expected cost C defined by

$$C = IC + AC,$$

with

$$E\Pi = P - C,$$

where IC and AC are respectively the expected insurance coverage and the expected audit cost.[27]

Insurance coverage is paid to the policyholders who actually experience a loss and to the opportunists who fraudulently report a loss and are not audited. We have

$$IC = t[\delta + \alpha\sigma(1-\delta)(1-p)] \tag{33}$$

$$AC = pc[\delta + \alpha\sigma(1-\delta)] \tag{34}$$

[27]For the sake of simplicity, we assume that no award is paid to the insurer when an opportunist is caught cheating. The fine B is entirely paid to the government.

As in the previous sections, we assume that the insurer can commit to his audit policy which means that he has a Stackelberg advantage in the audit game: the audit probability p is chosen to minimize C given the reaction function of opportunists. Since in the next section we want to contrast such an equilibrium with a situation where the insurer cannot commit to its audit policy, we refer to this commitment equilibrium with the upper index c. Let $\alpha^c(t,P,\sigma), p^c(t,P,\sigma)$ and $C^c(t,P,\sigma)$ be respectively the equilibrium strategies of opportunists and insurers and the equilibrium expected cost in an audit game (q,P,σ) under commitment to audit policy. Proposition 9 characterizes these functions.

Proposition 9 *Under commitment to audit policy, the equilibrium of an audit game (t,P,σ) is characterized by*

$$\begin{aligned} p^c(t,P,\sigma) &= 0 \text{ and } \alpha^c(t,P,\sigma)=1 \text{ if } c > c_0(t,P,\sigma), \\ p^c(t,P,\sigma) &= \widetilde{p}(q,P) \text{ and } \alpha^c(t,P,\sigma)=0 \text{ if } c \leq c_0(t,P,\sigma), \\ C^c(t,P,\sigma) &= \min\{t[\delta+\sigma(1-\delta)], \delta[t+\widetilde{p}(t,P)c]\}, \end{aligned}$$

where

$$c_0(t,P,\sigma) = \frac{(1-\delta)\sigma t}{\sigma\widetilde{p}(t,P)}.$$

The proof of Proposition 9 is straightforward. Only two strategies may be optimal for the insurer: either fully preventing fraud by auditing claims with probability $p=\widetilde{p}(t,P)$ which gives $\alpha=0$[28] or abstaining from any audit $(p=0)$ which gives $\alpha=1$. The optimal audit strategy is chosen so as to maximize C. Using (33) and (34) gives the result. Proposition 9 shows in particular that, given the contract (t,P), preventing fraud through an audit policy is optimal if the audit cost c is low enough and the proportion of opportunists σ is large enough.

We now consider a competitive insurance market with free entry, where insurers complete by offering policies. An adverse selection feature is brought in the model because the insurers cannot distinguish opportunists from honest policyholders. Following the approach of Wilson (1977), a market equilibrium is defined as a set of profitable contracts such that no insurer can offer another contract which remains profitable after the other insurers have withdrawn all non-profitable contracts in reaction to the offer. Picard (1996) characterizes the market equilibrium by assuming that honest individuals are

[28] $\alpha=0$ is an optimal strategy for opportunists when $p=\widetilde{p}(t,P)$ and it is the *only* optimal strategy if $p=\widetilde{p}(t,P)+\varepsilon, \varepsilon>0$.

uniformly distributed among the best contracts, likewise for opportunists. This assumption will be called **A**. Let[29]

$$(t^c, P^c) = \text{Arg Max}_{t,P}\{\delta U(W - L + t - P) + (1-\delta)U(W - P) \\ \text{s.t. } P \geq C^c(t, P, \theta)\}.$$

Proposition 10 *Under* **A**, (t^c, P^c) *is the unique market equilibrium when the insurers can commit to their audit policy.*

According to Proposition 10, a market equilibrium is defined by a unique contract (t^c, P^c) that maximizes the expected utility of honest policyholders under the constraint that opportunists cannot be set aside.[30] The arguments at work in the proof of Proposition 10 can be summarized as follows. Let us first note that all contracts offered at equilibrium are necessarily equivalent for honest customers, otherwise some equilibrium contracts would only attract opportunists. Given **A**, this would imply that $\alpha = 1$ is the equilibrium strategy of opportunists for such contract and these contracts could not be profitable. Equilibrium contracts are also equivalent for opportunists. Assume *a contrario* that opportunists concentrate on a subset of equilibrium contracts. For these contracts, the proportion of opportunists is larger than θ and honest individuals prefer (t^c, P^c) to these contracts. A contract $(t^c - \varepsilon, P^c), \varepsilon > 0$ would attract all honest individuals for ε small and would remain profitable even if opportunists finally also opt for this new contract. This contradicts the definition of a market equilibrium. Hence, for any contract (t, P) offered at the equilibrium, the insurers' participation constraint is $P \geq C^c(t, P, \theta)$. If (t^c, P^c) is not offered, then another contract could be proposed that would be strictly preferred by honest individuals and that would remain profitable whatever the reaction of opportunists. Hence (t^c, P^c) is the only possible market equilibrium. Another contract $(\widetilde{t}, \widetilde{P})$, offered in addition to (t^c, P^c) will be profitable if it attracts honest individuals only[31] and if $\widetilde{P} > \delta\widetilde{t}$. If $(\widetilde{t}, \widetilde{P})$ were offered, the insurers that go on offering (t^c, P^c) loose money. Indeed in such a case we necessarily have $\alpha^c(t^c, P^c, \widetilde{\sigma}) = 1$ where $\widetilde{\sigma}$ is the proportion of opportunists in the population of insureds who still choose (t^c, P^c) after $(\widetilde{t}, \widetilde{P})$ has been offered with $\widetilde{\sigma} > \theta$.[32]

[29] We assume that (t^c, P^c) is a singleton

[30] Proposition 10 shows that a pooling contract is offered at equilibrium: there does not exist any separating equilibrium where honest and opportunist individuals would choose different contracts. This result is also obtain by Boyer (1999) in a similar framework.

[31] Opportunists cannot benefit from separating and (t^c, P^c) is the best pooling contract for honest individuals.

[32] We have $\widetilde{\sigma} = 1$ if all honest policyholders choose $(\widetilde{t}, \widetilde{P})$ and $\widetilde{\sigma} = \frac{2\theta}{\theta+1}$ if $(\widetilde{t}, \widetilde{P})$ and (t^c, P^c) are equivalent for honest policyholders.

We then have

$$\begin{aligned} C^c(t^c, P^c, \sigma^c) &= t^c[\delta + \widetilde{\sigma}(1-\delta)] \\ &> t^c[\delta + \theta(1-\delta)] \geq C^c(t^c, P^c, \theta) = P^c, \end{aligned}$$

which proves that (t^c, P^c) becomes non-profitable. Hence (t^c, P^c) will be withdrawn and all individuals will turn toward the new contract $(\widetilde{t}, \widetilde{P})$. This new contract will show a deficit and it will not be offered, which establishes that (t^c, P^c) is the market equilibrium.

The market equilibrium is depicted in Figures 8 and 9. The perfect information market equilibrium is A with full insurance offered at fair premium.

Maximizing $EU = \delta U(W - L - P + t) + (1-\delta)U(W - P)$ with respect to $t \geq 0, P \geq 0$ subject to $P = \delta[t + c\widetilde{p}(t, P)]$ gives $t = \widehat{t}$ and $P = \widehat{P}$ at point B. We denote η_B the expected utility at B and we assume $\eta_B > \overline{U}$, i.e., the origin of the axis is over the indifference curve that goes through B. This assumption is satisfied if the audit cost c is not too large. Maximizing EU with respect to $t \geq 0, P \geq 0$ subject to $P = t[\delta + (1-\delta)\theta]$ gives $t = \overline{t}$ and $P = \overline{P}$ at point C. We denote $\eta_C(\theta)$the expected utility at C, with $\eta'_C(\theta) < 0$. Let $\widehat{\theta} \in (0,1)$ such that $\eta_B = \eta_C(\widehat{\theta})$. When $\theta > \widehat{\theta}$, the market equilibrium is at B: the insurers audit claims with probability $\widetilde{p}(\widehat{t}, \widehat{P})$ and the opportunists are detered from defrauding. When $\theta < \widehat{\theta}$, the market equilibrium is at C: the insurers do not audit claims because the proportion of opportunists is too small for verifying claims to be profitable and the opportunists systematically defraud. Hence, when $\theta < \widehat{\theta}$, there is fraud at equilibrium.

Figure 8

Figure 9

Here, we have assumed that the proportion of opportunistic individuals in the population is exogenously given. Note however that moral standards may be affected by the perception of insurers'honesty and also by beliefs about the prevalence of fraud among policyholders.[33] It has been widely documented in the business ethics literature that insurance defrauders often do not perceive insurance claim padding as an unethical behavior and even tend to practice some kind of self-justification. In particular, a common view

[33]Poverty may also affect morality. In particular, moral standards may decrease when the economic situation worsens. Dionne and Wang (2011) analyze the empirical relationship between opportunistic fraud and the business cycle in the Taiwan automobile theft insurance market. They show that fraud is stimulated during periods of recession and mitigated during periods of expansion.

among consumers holds that insurance fraud would just be the rational response to the unfair behavior of insurance companies. Tennyson (1997,2002) emphasizes that the psychological attitude toward insurance fraud is related to the perception of the fairness of insurance firms by policyholders. She shows that negative perceptions of insurance institutions are related to attitudes toward filing exaggerated claims. For instance, Tennyson (2002) shows that consumers who are not confident of the financial stability of their insurer and those who find auto insurance premiums to be burdensomely high are more likely than others to find fraud acceptable.Thus, consumers would tend to rationalize and justify their fraudulent claims through their negative perceptions of insurance companies.[34]

Fukukawa *et al.* (2007) substantiate this approach of the psychology of insurance defrauders.[35] They use a questionnaire to examine the factors that influence the decision-making of "aberrant consumer behaviors" (ACB) such as exagerating an insurance claim, but also changing a price tag, returning a stained suit, copying software from a friend and taking a quality towel from an hotel. Four factors emerged from a Principle Component Analysis, with among them the perception of unfairness relating to business practice[36]. Fukukawa *et a*l. (2007) show that the perceived unfairness factor is dominant in characterizing the occurence of the scenario where individuals exaggerate claims and that its effect on insurance fraud is significantly larger than on the other aberrant behavior scenarios.[37] Likewise, individuals'moral standards may depend on their perception of ethics heterogeneity : a policyholder may choose to be honest if he thinks this is the standard behavior in the society around him, but he may start cheating if he thinks "everybody does it". Such perceptions of social ethical standards would affect the proportion θ of opportunistic policyholders.

[34]See also Dean (2004) on the perception of the ethicality of insurance claim fraud.

[35]See also Strutton *et al.* (1994) on how consumers may justify inappropriate behavior in market settings.

[36]The *Perceived Unfairness* factor is comprised of items related to the perception of unfair business practice, for instance because the insurer is overcharging, or because ACB is nothing but retaliation against some inadequate practice or because of weak business performance. Other factors are labeled *Evaluation* (loading variables relating to the easiness to engage in ACB or to the general attitude toward ACB), *Social Participation* (with variables representing the social external encouragement to ACB) and *Consequence* (measuring the extent to which the outcomes of ACB are seen as beneficial or harmful).

[37]See Bourgeon and Picard (2012) for a model where policyholder's moral standards depend on the attitude of insurers who may nitpick claims and sometimes deny them if possible.

7 The credibility issue

In a situation where there are many opportunist policyholders, it is essential for insurers to credibly announce that a tough monitoring policy will be enforced, with a high probability of claim verification and a high level of scrutinity for suspected fraud. In the model introduced in the previous section, this was reached by announcing that claims are audited with probability $\widetilde{p}(t,P)$. However, since auditing is costly to the insurer, a commitment to such a tough audit policy may not be credible.

In the absence of commitment, i.e., when the insurer has no Stackelberg advantage in the audit game, the auditing strategy of the insurer is constrained to be a best response to opportunists' fraud strategy, in a way similar to tax compliance games[38] studied by Graetz *et al.* (1986) and Melumad and Mookherjee (1989).[39] In the model introduced in the previous section, under no commitment to audit policy, the outcome of an audit game (t,P,σ) corresponds to a perfect Bayesian equilibrium, where: (a) the fraud strategy is optimal for an opportunist given the audit policy, (b) the audit policy is optimal for the insurer given beliefs about the probability of a claim to be fraudulent, (c) the insurer's beliefs are obtained from the probability of loss and opportunists strategy using Bayes' rule.

Let $\alpha^n(t,P,\sigma)$ and $p^n(t,P,\sigma)$ be the equilibrium strategy of opportunists and of insurers respectively, in an audit game in the absence of commitment to an audit policy and let $C^n(t,P,\sigma)$ be the corresponding expected cost.

Proposition 11 *Without commitment to an audit policy, the equilibrium of an audit game* (t,P,σ) *is characterized by*[40]

$$\begin{aligned}
p^n(t,P,\sigma) &= 0 \text{ and } \alpha^n(t,P,\sigma)=1 \text{ if } c>c_1(t,\sigma),\\
p^n(t,P,\sigma) &= \widetilde{p}(t,P) \text{ and } \alpha^n(t,P,\sigma)=\frac{\delta c}{\sigma(1-\delta)(t-c)} \text{ if } c>c_1(t,\sigma),\\
C^n(t,\sigma) &= \min\left\{t[\delta+\sigma(1-\delta)],\frac{\delta t^2}{t-c}\right\},
\end{aligned}$$

where

$$c_1(t,\sigma)=\frac{\sigma(1-\sigma)t}{\sigma(1-\sigma)+\delta}.$$

[38] See Andreoni *et al.* (1998) for a survey on tax compliance.

[39] Cummins and Tennyson (1994) analyze liability claims fraud within a model without Stackelberg advantage for insurers: each insurer chooses his fraud control level to minimize the costs induced by fraudulent claims.

[40] We assume $t>c$ and we neglect the case $c=c_1(t,\sigma)$. See Picard (1996) for details.

The proof of Proposition 11 may be sketched as follows. Let π be the probability for a claim to be fraudulent. Bayes' rule gives

$$\pi = \frac{\alpha\sigma(1-\delta)}{\alpha\sigma(1-\delta)+\delta}. \tag{35}$$

Once a policyholder puts in a claim, the (conditional) insurer's expected cost is

$$\overline{C} = p[c + (1-\pi)t] + (1-p)t. \tag{36}$$

The equilibrium audit policy minimizes $\overline{C}$ with respect to p which gives

$$\left.\begin{array}{cc} p = 0 & \text{if } \pi t < c, \\ p \in [0,1] & \text{if } \pi t = c, \\ p = 1 & \text{if } \pi t > c. \end{array}\right\} \tag{37}$$

The equilibrium of the no-commitment audit game is a solution (α, p, π) to (32), (35) and (37). Let us compare Proposition 11 to Proposition 9. At a no-commitment equilibrium, there is always some degree of fraud: $\alpha = 0$ cannot be an equilibrium strategy since any audit policy that totally prevents fraud is not credible. Furthermore, we have $c_1(t,\sigma) < c_0(t,P,\sigma)$ for all t, P, σ which means that the optimal audit strategy $p = \widetilde{p}(t,P,\sigma)$ that discourages fraud is optimal for a larger set of contracts in the commitment game than in the no-commitment game. Lastly, we have $C^n(t,\sigma) \geq C^c(t,P,\sigma)$ with a strong inequality when the no-commitment game involves $p > 0$ at equilibrium. Indeed, at a no-commitment equilibrium, there must be some degree of fraud for an audit policy to be credible which increases insurance expected cost.[41]

The analysis of market equilibrium follows the name logic as in the commitment case. Let

$$\begin{aligned}(t^n, P^n) = \quad & \text{Arg Max}_{t,P}\{\delta U(W-L+t-P) + (1-\delta)U(W-P) \\ & \text{s.t. } P \geq C^n(t,P,\theta)\end{aligned}$$

be the pooling contract that maximizes the expected utility of honest policyholders.[42]

Proposition 12 *Under* $\mathbf{A}$, (t^n, P^n) *is the unique market equilibrium when the insurers cannot commit to their audit policy.*

[41] As shown by Boyer (1999), when the probability of auditing is strictly positive at equilibrium (which occurs when θ is large enough), then the amount of fraud $(1-\delta)\theta\alpha^n(t^n, P^n, \theta) = \delta c/(t^n - c)$ does not depend on θ. Note that t^n does not (locally) depend on θ when $c < c_1(t^n, \theta)$.

[42] We assume that (t^n, P^n) is a singleton.

The expected utility of honest policyholders is higher at the commitment equilibrium than at the no-commitment equilibrium. To highlight the welfare costs of the no-commitment constraint, let us focus attention on the case where θ is sufficiently large so that, in the absence of claims' verification, honest customers would prefer not to take out an insurance policy than to pay high premiums that cover the cost of systematic fraud by opportunists. This means that point C is at the origin of the axis in Figures 8 and 9, which occurs if $\theta \geq \theta^*$, with

$$\theta^* = \frac{\delta[U'(W-L) - U'(W)]}{\delta U'(W-L) + (1-\delta)U'(W)} \in (0,1).$$

In Figure 10, the commitment equilibrium is at point B (i.e., $\theta < \theta^*$) and the no-commitment equilibrium is at the origin of the axis: the market shuts down completely at $t = t^n = 0$.[43]

Figure 10

Hence, besides the inevitable market inefficiency induced by the cost of auditing (i.e., going from A to B in Figure 10), the inability of insurers to commit to an audit policy induces an additional welfare loss (from B to 0). How can this particular inefficiency be overcome? Two solutions have been put forward in the literature. A first solution, developped by Melumad and Mookherjee (1989) in the case of income tax audits, is to delegate authority over an audit policy to an independant agent in charge of investigating claims. An incentive contract offered by the insurer to the investigator could induce a tough monitoring strategy, and precommitment effects would be obtained by publicly announcing that such incentives have been given to the investigator. Secondly, Picard (1996) shows that transferring audit costs to a budget balanced common agency may help to solve the commitment problem. The common agency takes charge of part of the audit expenditures decided by insurers and is financed by lump-sum participation fees. This mechanism mitigates the commitment problem and may even settle it completely if there is no asymmetric information between the agency and the insurers about audit costs. Thirdly, Krawczyk (2009) shows that putting the

[43] It can be shown that $t^n > L$ when there is some audit at equilibrium, that is when $\theta > \theta^*$. Boyer (2004) establishes this result in a slightly different model. Intuitively, increasing t over L maintains the audit incentives at the right level for a lower fraud rate π, because we should have $\pi t = c$ for $p = \widetilde{p}(t,P) \in (0,1)$ to be an optimal choice of insurers. In the neighbourhood of $t = L$, an increase in t only induces second-order risk-sharing effects, and ultimately that will be favorable to the insured.

insurer-policyholder interaction in a dynamic context reduces the intensity of the commitment problem. More specifically, he nests insurer-policyholder encounters into a supergame with a sequence of customers. Using the folk theorem for repeated games with many short-lived agents (see Fudenberg *et al.*,1990), he shows that the capacity of an insurer to develop its reputation for "toughness" and deter fraud depends on the observability of its auditing strategy. Under full observability of the mixed auditing strategy, fraud can be fully detered, provided the insurers' discount factor is large enough, i.e. they are sufficiently patient. More realistically, if policyholders base their decisions on sampling information from the past period, then only partial efficiency gains are possible and the larger the size of the sample of observed insurer-policyholder interactions, the lower the frequency and thus the cost of fraud. Hence, signalling claims monitoring effort to policyholders should be part and parcel of the struggle against insurance fraud.

8 Using fraud signals

When there is a risk of fraud, it is in the interest of insurers to use signals on agents' losses when deciding whether a costly verification should be performed. This leads us to make a connection between optimal auditing and scoring techniques.

We will start with the simple case where the insurer perceives a binary signal $s \in \{s_1, s_2\}$ when a policyholder files a claim. The signal s is observed by the insurer and it cannot be controlled by defrauders. Let q_i^f and q_i^n be, respectively, the probability of $s = s_i$ when the claim is fraudulent (i.e., when no loss occurred) and when it corresponds to a true loss, with $0 < q_2^n < q_2^f$ and $q_1^n + q_2^n = q_1^f + q_2^f = 1$. Thus, we assume that s_2 is more frequently observed when the claim is fraudulent than when it corresponds to a true loss : s_2 may be interpreted as a fraud signal that should make the insurer more suspicious. The decision to audit can now be conditioned on the perceived fraud signal. Let us first assume that the insurer can commit to its auditing strategy. $\widehat{p}_i \in [0,1]$ denotes the audit probability when signal s_i is perceived, with $\widehat{p} = (\widehat{p}_1, \widehat{p}_2)$. $\widetilde{p}(t,P)$ still denotes the audit probability that deters opportunistic individuals from filing fraudulent claims. If $q_2^f \geq \widetilde{p}(t,P)$, then fraud is detered if $\widehat{p}_2 \geq \widetilde{p}(t,P)/q_2^f$ and $\widehat{p}_1 = 0$. If $q_2^f < \widetilde{p}(t,P)$, the fraud is detered if $\widehat{p}_2 = 1$ and $\widehat{p}_1 = [\widetilde{p}(t,P) - q_2^f]/(1-q_2^f) < 1$. In other words, we here assume that the insurer's auditing strategy prioritizes the claims with signal s_2. If auditing these claims with probability one is not enough for fraud to be deterred, then a proportion of the claims with signal s_1 are also audited. We will check later that such a strategy is optimal. For the sake of brevity, let us

consider here the case where the optimal constract is such that $q_2^f > \widetilde{p}(t,P)$. Expected insurance cost IC and expected audit cost AC are now written as

$$\begin{aligned} IC &= t[\delta + \alpha\sigma(1-\delta)(1-\widehat{p}_2 q^f)], \\ AC &= \widehat{p}_2 c[q^n\delta + \alpha\sigma(1-\delta)q^f], \end{aligned}$$

with unchanged definitions of $\delta, \alpha, \sigma, c$ and t. As in Section 6, the insurer may decide either to deter fraud by opportunistic individuals - he would choose $\widehat{p}_1 = 0, \widehat{p}_2 = \widetilde{p}(t,P)/q_2^f \in (0,1)$ - or not ($\widehat{p}_1 = \widehat{p}_2 = 0$). When fraud is detered, we have $\alpha = 0$ and the cost of a claim is

$$C = \delta(t + \widehat{p}_2 c q_2^n) = \delta[t + \widetilde{p}(t,P)c\frac{q_2^n}{q_2^f}].$$

Thus, under commitment to audit policy with fraud signal s, the expected cost in an audit game (t,P,σ) is

$$\widehat{C}^c(t,P,\sigma) = \min\{t[\delta + \sigma(1-\delta)], \delta[t + \widetilde{p}(t,P)c\frac{q_2^n}{q_2^f}]\}.$$

$q_2^n/q_2^f < 1$ implies $\widehat{C}^c(t,P,\sigma) \leq C^c(t,P,\sigma)$, with a strong inequality when detering fraud is optimal. We deduce that conditioning auditing on the fraud signal reduces the claims cost, and ultimately it increases the expected utility of honest individuals for the optimal contract[44].

The previous reasoning may easily be extended to the more general case where the insurer perceives a signal $s \in \{s_0, s_1, ..., s_\ell\}$ with $\ell \geq 2$, following Dionne *et al.* (2009).[45] Let q_i^f and q_i^n be respectively the probability of the signal vector s taking on value s_i when the claim is fraudulent and when it corresponds to a a true loss, with $\sum_{i=1}^{\ell} q_i^f = \sum_{i=1}^{\ell} q_i^n = 1$. Without loss of generality, we assume $q_i^n > 0$ for all i and we rank the possible signals in such a way that[46]

$$\frac{q_1^f}{q_1^n} < \frac{q_2^f}{q_2^n} < ... < \frac{q_\ell^f}{q_\ell^n}.$$

[44] As before, the optimal contract maximizes the expected utility of honest policyholders under the constraint $P \geq \widehat{C}^c(t,P,\theta)$, where θ still denotes the proportion of opportunist individuals in the population. If the optimal contract without fraud signal is such that $\delta[t + \widetilde{p}(t,P)c\frac{q_2^n}{q_2^f}] < t[\delta + \theta(1-\delta)] < \delta[t + \widetilde{p}(t,P)c]$, then auditing claims is optimal only if the insurer can condition his decision on the fraud signal.

[45] As in Dionne *et al.* (2009), s may be a k-dimensional signal, with k the number of fraud indicators (or red flags) observed by the insurer. Fraud indicators cannot be controlled by defrauders and they may make the insurer more suspicious about fraud. For instance, when all indicators are binary, then $\ell = 2^k$ and s may be written as a vector of dimension k with components 0 or 1 : component j is equal to 1 when indicator j is "on", and it is equal to 0 when it is "off".

[46] Of course if $q_i^n = 0$ and $q_i^f > 0$, then it is optimal to trigger an audit when $s = s_i$ because the claim is definitely fraudulent in that case.

With this ranking, we can interpret $i = 1, ..., \ell$ as an index of fraud suspicion. Indeed, assume that the insurer consider that a claim may be fraudulent with (*ex ante*) probability π^a. Then, using Bayes law allows us to write the probability of fraud (fraud score) conditional on signal s_i as

$$\Pr(Fraud\,|s_i) = \frac{q_i^f \pi^a}{q_i^f \pi^a + q_i^n(1-\pi^a)},$$

which is increasing with i. Thus, as index i increases so does the probability of fraud.[47]

Now the insurers auditing strategy is written as $\widehat{p} = (\widehat{p}_1, \widehat{p}_2, ...\widehat{p}_\ell)$ where $\widehat{p}_i \in [0,1]$ denotes the audit probability when signal s_i is perceived. Fraudulent and non-fraudulent claims are audited with probability $\sum_{i=1}^{\ell} q_i^f \widehat{p}_i$ and $\sum_{i=1}^{\ell} q_i^n \widehat{p}_i$ respectively. Opportunistic individuals are detered from defrauding if $\sum_{i=1}^{\ell} q_i^f \widehat{p}_i \geq \widetilde{p}(t,P)$ and in that case the expected cost of a claim is written as the sum of the indemnity t and the expected audit cost $c\sum_{i=1}^{\ell} q_i^n \widehat{p}_i$. Thus, the optimal fraud detering audit strategy $\widehat{p} = (\widehat{p}_1, \widehat{p}_2, ...\widehat{p}_\ell)$ minimizes the expected cost of claims

$$t + c\sum_{i=1}^{\ell} q_i^n \widehat{p}_i,$$

subject to

$$\sum_{i=1}^{\ell} q_i^f \widehat{p}_i \geq \widetilde{p}(t,P),$$
$$0 \leq \widehat{p}_i \leq 1 \text{ for all } i = 1, ..., \ell.$$

This is a simple linear programming problem, whose optimal solution is characterized in the following Proposition :

Proposition 13 *An optimal auditing strategy is such that*

$$\widehat{p}_i = 0 \text{ if } i < i^*,$$
$$\widehat{p}_i \in (0,1] \text{ if } i = i^*,$$
$$\widehat{p}_i = 1 \text{ if } i > i^*,$$

[47] In the present model, insurers fully deter fraud when they can commit to their auditing strategy and the proportion of opportunist individuals is large enough. This is no longer true when there is a continuum of types for individuals. Dionne *et al.* (2009) consider such a model, with a continuum of individuals and morale costs that may be more or less important. In their model, there is a positive rate of fraud even if insurers can commit to their audit strategy. π^a would then correspond to the equilibrium fraud rate, which is positive, but lower than the equilibrium fraud rate under the no-commitment hypothesis.

where $i^* \in \{1, ..., \ell\}$*, and when fraud is detered the audit probability is*

$$p^c(t,P) \equiv \sum_{i=1}^{\ell} q_i^n \widehat{p}_i < \widetilde{p}(t,P).$$

Proposition 13 says that an optimal verification strategy consists in auditing claims when the suspicion index i exceeds the critical threshold i^*. Thus, the insurer plays a "red flags strategy" : for some signals - those with $i > i^*$ - claims are systematically audited, wheras there is no audit when $i < i^*$ and audit is random when $i = i^*$.[48] Choosing $\widehat{p}_i = \widetilde{p}(t,P)$ for all $i = 1, ..., \ell$ is a suboptimal fraud detering strategy. Thus, using fraud signals allow to audit a smaller fraction of claims while detering fraud. The expected cost per policyholder is

$$\widehat{C}^c(t,P,\sigma) = \min\{t[\delta + \sigma(1-\delta)], \delta[t + cp^c(t,P)]\}$$

with $p^c(t,P) < \widetilde{p}(t,P)$. Thus, we have $\widehat{C}^c(t,P,\sigma) \leq C^c(t,P,\sigma)$, with a strong inequality when it is optimal to deter fraud, which shows that insurers can reduce the cost of claims by triggering audit on the basis of fraud signals.

Let us turn to the case where insurers cannot commit to their auditing strategy, and once again let us start with a binary signal $s \in \{s_1, s_2\}$ with $0 < q_2^n < q_2^f$.[49] The insurers' auditing strategy should then be the best response to the opportunistic policyholders' fraud strategy. α still denotes the fraud rate of opportunistic individuals and the proportion of fraudulent claims π is still given by (35). Let us focus once again on the case where $q_2^f > \widetilde{p}(t,P)$, so that it is possible to deter fraud by auditing claims under signal s_2, with $\widehat{p}_1 = 0$. Assume first $\widehat{p}_2 > 0$. The expected cost of a claim is

$$\overline{C} = (1-\pi)\left(t + cq_2^n \widehat{p}_2\right) + \pi\left(t - (t-c)q_2^f \widehat{p}_2\right),$$

which extends (36) to the case where the audit probability differs between fraudulent claims and non-fraudulent claims. As in Section 7, there cannot exist an equilibrium where fraud would be fully detered : indeed $\alpha = 0$ would give $\pi = 0$ and $\widehat{p}_2 = 0$ and then $\alpha = 1$ would be an optimal fraud strategy of opportunistic individuals, hence a contradiction. When $\alpha = 1$, we necessarily have $q_2^f \widehat{p}_2 \leq \widetilde{p}(t,P)$, and (35) then gives

$$\pi = \frac{\sigma(1-\delta)}{\sigma(1-\delta) + \delta} \equiv \overline{\pi}.$$

[48] If $\ell = 2$ and detering fraud is optimal, then we have $i^* = 2$ if $q_2^f \geq \widetilde{p}(t,P)$ and $i^* = 1$ if $q_2^f < \widetilde{p}(t,P)$.

[49] This case has been studied by Schiller (2003).

Assume $q_2^f/q_2^n > c(1-\overline{\pi})/\overline{\pi}(t-c)$. In that case, minimizing $\overline{C}$ with respect to $\widehat{p}_2 \in [0,1]$, with $\pi = \overline{\pi}$, would give $\widehat{p}_2 = 1$, and thus $q_2^f \widehat{p}_2 > \widetilde{p}(t,P)$ which contradicts $\alpha = 1$. Thus $\alpha \in (0,1)$ is the only possible case, which implies $\widehat{p}_2 = \widetilde{p}(t,P)/q_2^f \in (0,1)$ and $\pi \in (0,\overline{\pi})$. For $\overline{C}$ to be minimized at $\widehat{p}_2 = \widetilde{p}(t,P)/q_2^f \in (0,1)$, we need to have

$$\pi = \frac{cq_2^n}{cq_2^n + (t-c)q_2^f},$$

which implies $\overline{C} = t$ and

$$\begin{aligned} C &= \frac{\delta \overline{C}}{1-\pi} \\ &= \frac{\delta t[cq_2^n + (t-c)q_2^f]}{(t-c)q_2^f}. \end{aligned}$$

When $\widehat{p}_2 = 0$, we have $C = t[\delta + \sigma(1-\delta)]$. Thus, when the insurer cannot commit to its audit strategy, the expected cost in an audit game (t,P,σ) is

$$\widehat{C}^n(t,\sigma) = \min\{t[\delta + \sigma(1-\delta)], \frac{\delta t[cq_2^n + (t-c)q_2^f]}{(t-c)q_2^f}\}.$$

Using $q_2^n < q_2^f$ yields $\widehat{C}^n(t,\sigma) \leq C^n(t,\sigma)$, with a strong inequality when it is optimal to audit claims with positive probability. Thus, conditioning audit on fraud signals reduces the cost of claims even if the insurer cannot commit to its verification strategy.

If the insurer perceives a signal $s \in \{s_0, s_1, ..., s_\ell\}$ with q_i^f/q_i^n increasing in i, then the expected cost of a claim is

$$\overline{C} = (1-\pi)\left(t + c\sum_{i=1}^{\ell} q_i^n \widehat{p}_i\right) + \pi\left(t - (t-c)\sum_{i=1}^{\ell} q_i^f \widehat{p}_i\right),$$

The optimal auditing strategy minimizes $\overline{C}$ with respect to $\widehat{p} = (\widehat{p}_1, \widehat{p}_2, ...\widehat{p}_\ell)$ subject to $0 \leq \widehat{p}_i \leq 1$ for all $i = 1, ..., \ell$. We deduce that $\widehat{p}_i = 0$ for all i if $\pi = 0$. If $\pi > 0$, then we have :

$$\begin{aligned} &\widehat{p}_i = 0 \ \text{ if } \ \frac{q_i^f}{q_i^n} < \frac{c(1-\pi)}{\pi(t-c)}, \\ &\widehat{p}_i \in [0,1] \ \text{ if } \frac{q_i^f}{q_i^n} = \frac{c(1-\pi)}{\pi(t-c)}, \\ &\widehat{p}_i = 1 \text{ if } \frac{q_i^f}{q_i^n} > \frac{c(1-\pi)}{\pi(t-c)}. \end{aligned}$$

Since q_i^f/q_i^n is increasing in i, we deduce that the characterization given in Proposition 13 is also valid in the no-commitment case. Audit is triggered when a suspicion index i^* is reached, where i^* is the smallest index i such that $q_i^f/q_i^n \geq c(1-\pi)/\pi(t-c)$. As in the case of a binary signal, there cannot exist an equilibrium where fraud would be fully detered. Thus, we have $\alpha > 0$, and there is some fraud at equilibrium. The *ex ante* fraud probability π^a coincides with the proportion of fraudulent claims π. The larger the suspicion index, the larger the fraud score $\Pr(Fraud\,|s_i) = q_i^f\pi/(q_i^f\pi + q_i^n(1-\pi)$, and audit should be triggered when this fraud score is larger than c/t.

9 Some indirect effects of insurance contracts on fraud

As mentioned in Section 6, the intensity of insurance fraud may depend on the perception of unfair behavior on the part of insurance companies, in relation to some stipulations of insurance contracts. For instance Dionne and Gagné (2001) have shown with data from Québec that the amount of the deductible in automobile insurance is a significant determinant of the reported loss, at least when no other vehicle is involved in the accident, and thus when the presence of witnesses is less likely. This suggests that the larger the deductible, the larger the propensity of drivers to file fraudulent claims. Although a deductible is a clause of the insurance contract that cannot be interpreted as a bad faith attitude of the insurer, the result of Dionne and Gagné sustains the idea that the larger the part of an accident cost born by a policyholder, the larger the incentives he or she feels to defraud. In the same vein, the results of an experimental study by Miyazaki (2009) shows that higher deductibles result in weaker perception that claim padding is an unethical behavior, with the conclusion that the results indicate "some degree of perceived corporate unfairness, wherein consumers feel that the imbalance in favor of the firm has to be balanced by awarding the claimant a higher dollar amount".

Independently of induced effects on moral standards, some contractual insurance provisions may prompt dishonest policyholders to defraud and in that case, the risk of fraud should be taken into account in the design of optimal insurance policies. An example is provided by Dionne and Gagné (2002) through their analysis of replacement cost endorsement in automobile insurance. A replacement cost endorsement allows the policyholder to get a new car in the case of a theft or if the car has been totally destroyed in a road accident, usually if the theft or the collision occured in the first two years of

ownership of a new car. Such endorsements increase the protection of the insureds against depreciation, but they also increase the incentives to defraud, for instance by framing a fraudulent theft. Note that, in an adverse selection setting, an individual may choose to include a replacement cost endorsement in his coverage because he knows he will be more at risk. Furthermore, individuals may decide to drive less carefully or pay less attention to the risk of theft when their coverage is complete than when it is partial and thus, replacement cost endorsements may increase the insurance losses because of moral hazard. Thus, the fact that policyholders with a replacement cost endorsement have more frequent accidents or thefts may be the consequence of fraud, but it may also reflect adverse selection or moral hazard. Dionne and Gagné (2002) use data from Québec to disentangle these three effects. They show that holders of car insurance policies with a replacement cost endorsement have a higher probability of theft near the end of this additional protection (which usually lasts for two years after the acquisition of a new car). Their statistical tests rule out (ex ante) moral hazard and adverse selection[50] and they interpret their result as the effect of replacement cost endorsement on the propensity to defraud.

Another example of induced effects of contracts on the propensity to defraud occurs in the case of corporate property insurance. Following the accidental destruction of productive assets (e.g., buildings, plant, inventories), a firm must decide whether to restore those assets to their previous state and the contractual indemnity usually differs according to whether there is restoration or insurance payment. In such a setting, Bourgeon and Picard (1999) characterize the optimal corporate fire insurance contract when the insured firm has private information about the economic value of the damaged productive assets. They show that the indemnity should be larger in case of restoration than when the firm receives insurance money, but there should be partial coverage as well when restoration is chosen. The structure of indemnity payments is chosen to minimize the rent the firms enjoy when the (unverifiable) economic losses are smaller than the insurance payment, but also to prevent the firm from inefficient restoration (i.e., restoration when the economic value of the damaged capital is low). In this context, fraud may take the form of arson : arson may be decided on by dishonest firms that are in a position to set unprofitable equipment on fire to obtain insurance money. The possibility of arson is an additional motive for lowering insur-

[50]Moral hazard is ruled out because there is no significant effect of replacement cost endorsements on partial thefts (i.e., thefts where only a part of the car is stolen : hubcaps, wheels, radio,etc.) although the same self-protection activities affect the claims distribution of total and partial thefts. Dionne and Gagné (2002) also rule out adverse selection because the effect is significant for only one year of ownership and not for all years.

ance money under the restoration indemnity. Bourgeon and Picard (1999) show that, because of the risk of arson, the insurer may be led not to offer any insurance money to the firm but only to reimburse restoration costs.[51]

Experience rating, and particularly bonus-malus rules, may alleviate the propensity of opportunistic policyholders to defraud. Bonus-malus pricing in automobile insurance is usually viewed either as a risk type learning process under adverse selection or as an incentive device under moral hazard. However bonus-malus also affects the propensity to defraud when the mere fact of filing a claim, be it fraudulent or not, leads the insurer to charge higher rates in future periods. Bonus-malus rules may thus be of aid for reducing insurance fraud. This intuition has been developed by Moreno *et al.* (2006). The key ingredient of their model is the intertemporal choice of policyholders. For simplicity, they assume that at period t individuals only care about their utility during the current and following period t and $t+1$, and not about subsequent periods $t+2, ...$ (which is an extreme form of non-exponential discounting) and at each period a loss of a given size may occur. An opportunity for fraud exists when no loss occurs and the policyholder may fraudulently report a loss to the insurer. The insurer does not audit claims, but simply pays out on any filed claim. However the period following a claim, the insurer adjusts the premium according to whether or not a claim was filed. Thus, filing a fraudulent claim results in a present benefit to the policyholder at the cost of a higher future premium. Moreno *et al.* (2006) show that this trade-off may tip in favor of honesty, in the cases of a monopolist insurer and of a perfectly competitive markets, and they exhibit a condition for the bonus-malus anti-fraud mechanism to Pareto dominate the audit mechanism.

10 Collusion with agents

In many cases, insurance fraud goes through collusion between policyholders and a third party. For instance, collusion with auto mechanics, physicians or attorneys is a channel through which an opportunist policyholder may manage to falsify his claims. Falsification costs—taken as exogenous in the sections 3 and 4—then are the outcome of hidden agreement between policyholders and such agents.

[51] Bourgeon and Picard (1999) also consider stochastic mechanisms in which the restoration of damaged assets is an option given by the insurance contract to the insurer but not always carried out at equilibrium. The (randomly exercised) restoration option is used as a screening device : larger indemnity payments require larger probabilities of restoration, which prevents firms with low economic losses from building up their claims.

In this section, we focus on collusion between policyholders and agents in charge of marketing insurance contracts. We also consider another type of fraud, namely the fact that policyholders may lie or not disclose relevant information when they take out their policy.[52] We will assume that the agent observes a number of characteristics of the customer that allow him to estimate correctly the risks and to price the policy. These characteristics cannot be verified by the insurer. Agents also provide promotional services that affect the demand for the policies offered by the insurer but promotional effort cannot either be verified by the insurer.[53] The insurer only observes two signals of his agent's activity, namely net premiums written and indemnity payments.

The key element we want to focus on is the fact that agents may be willing to offer unduly advantageous contracts to some policyholders in order to compensate low promotional efforts. This possibility should lead the insurer to condition his agents' commissions at the same time on cashed premiums and on indemnity payments. Of course, the issue of how an insurer should provide incentives to his selling agents—be they exclusive or independant—is important independently of insurance fraud. However, in a situation where the insurer does not perfectly monitor his agents, there is some scope for collusion between agents and policyholders which facilitates insurance fraud. The agent may be aware of the fact that the customer tells lies or that he conceals relevant information but he overlooks this violation in order not to miss an opportunity to sell one more insurance policy. Hence, in such a case, the defrauder is in fact the policyholder-agent coalition itself. In what follows, we sketch a model that captures some consequences of insurance fraud through collusion between policyholders and agents.

Consider an insurance market with n risk-neutral firms of equal size. Each firm employs ℓ exclusive agents to sell insurance contracts.[54] Let e be the promotional effort expended by an agent. Let k be the loading factor used to price the policies written by the agent. For any customer, the agent is supposed to be able to correctly estimate the expected indemnity payments Et. Let $\widehat{k}$ be the loading factor decided upon by the insurer. Hence, if

[52] On this kind of fraud where insurers can (at some cost) verify the policyholders' types, see Dixit (2000), Dixit and Picard (2003) and Picard (2009).

[53] The choice of distribution system affects the cost to the insurers of elicitating additional promotional effort of their sales force. For instance, exclusive representation prevents the agents from diverting potential customers to other insurers who pay larger commissions. Likewise giving independent agents ownership of policy expirations provides incentives for agents to expend effort to attract and retain customers—see Kim *et al.* (1996).

[54] Modelling promotional effort in an independent agency system would be more complex since, in such a system, the agent's decisions are simultaneously affected by several insurers.

expected indemnity payments are truthfully reported by the selling agent to the insurer, the pricing rule should lead the agent to charge a premium $(1+\widehat{k})Et$. However, by misreporting expected indemnity payments, the agent is able to write policies with an actual loading factor lower than $\widehat{k}$. In what follows, e and k are the decision variables of the agent.

Let P and Q be respectively the aggregate premiums collected by a given agent and the aggregate indemnity payments made to his customers during a period of time. We assume

$$P = \frac{1}{n\ell}[g(e,k) + \varepsilon_1] \text{ with } g'_e > 0 \text{ and } g'_k < 0 \tag{38}$$

where ε_1 is an idiosyncratic random parameter that varies among agents, with $E\varepsilon_1 = 0$. ε_1 is unknown when the selling agent chooses e and k and cannot be observed by the insurer. Larger promotional efforts increase the amount of collected premiums. Furthermore, we assume that the elasticity of demand for coverage (in terms of expected insurance demand) with respect to loading $1+k$ is larger than one. Hence a higher loading factor—or, equivalently, less downward misreporting of expected insurance payments by the agent to the insurer—decreases the premiums cashed. Note that the coefficient $1/n\ell$ in (38) reflects the market share of each agent. We also have

$$Q = \frac{1}{n\ell}\left[h(e,k) + \frac{\varepsilon_1}{l+k} + \varepsilon_2\right] \tag{39}$$

where $h(e,k) \equiv g(e,k)/\ell + k$, with $h'_e > 0, g'_k < 0$ and where ε_2 is another idiosyncratic random parameter, uncorrelated with ε_1, such that $E\varepsilon_2 = 0$.

Let $\Psi(e)$ be the cost to the agent of providing promotional effort at level e, with $\Psi' > 0, \Psi'' > 0$. The agents are supposed to be risk-averse.

If insurers were able to monitor the promotional effort and to verify the expected indemnity payments of the policies written by their agents, they would be in position to choose e and k so as to maximize their expected profit written as

$$E\Pi = \ell[EP - EQ - EC]$$

where C denotes the commission paid to each agent. Under perfect information about the agent's behaviour it is optimal to pay fixed commissions so that net earnings $C - \Psi(e)$ are equal to a given reservation payment normalized at zero. We thus have $C = \Psi(e)$, which gives

$$E\Pi = \frac{1}{n}[g(e,k) - h(e,k)] - \ell\Psi(e) \tag{40}$$

Maximizing $E\Pi$ with respect to e and k gives the first best solution $e = e^*$ and $k = k^*$. A free entry perfect information equilibrium is defined by $E\Pi = 0$ which gives an endogenously determined number of firms $n = n^*$.

Assume now that the insurers do not observe the promotional effort expended by the agents. They can neither verify the expected indemnity payments associated with the policies written by their agent. Opportunist policyholders would like to purchase insurance priced at a loading factor lower than $\widehat{k}$ by not disclosing relevant information about the risks incurred to the insurer. It is assumed that this hidden information cannot be revealed to the insurer if an accident occurs. The agent observes the risks of the customers but he may choose not to report this information truthfully to the insurer in order to get larger sales commissions. The insurer may control the agent opportunism by conditioning his commissions both on cashed premiums and on indemnity payments. However, because of the uncertainty that affects premiums and losses, risk premiums will have to be paid to selling agents which will ultimately affect the firm's profitability.

Assume that the commission paid to an agent depends linearly on P and Q, i.e.

$$C = \alpha P - \beta Q + \gamma$$

Assume also that the agents' utility function V is quadratic, which allows us to write

$$EV = EC - \rho Var(C) - \Psi(e) \text{ with } \rho > 0$$

The agent's participation constraint $EV \geq 0$ is binding at the optimum, which gives

$$\begin{aligned} EC &= \rho Var(C) + \Psi(e) \\ &= \frac{\rho}{(n\ell)^2}\left[\alpha^2\sigma_1^2 + \beta^2\frac{(\sigma)_1^2}{(1+k)^2} + \beta^2\sigma_2^2\right] \end{aligned}$$

where $\sigma_1^2 = Var(\varepsilon_1)$ and $\sigma_2^2 = Var(\varepsilon_2)$. We obtain

$$E\Pi = \frac{1}{n}[g(e,k) - h(e,k)] - \ell\Psi(e) - \frac{\rho}{n^2\ell}\left[\alpha^2\sigma_1^2 + \beta^2\frac{\sigma_1^2}{(1+k)^2} + \beta^2\sigma_2^2\right]$$

The insurer maximizes $E\Pi$ with respect to $e \geq 0, k \geq 0, \alpha$ and β subject to the agent's incentive compatibility constraint

$$\begin{aligned} (e,k) \in \text{Arg Max}_{e',k'} EV = &\frac{\alpha}{n\ell}g(e',k') - \frac{\beta}{n\ell}h(e',k') + \gamma - \Psi(e') \\ &-\frac{\rho}{(n\ell)^2}\left[\alpha^2\sigma_1^2 + \beta^2\frac{\sigma_1^2}{(1+k')^2} + \beta^2\sigma_2^2\right] \end{aligned}$$

If there is some positive level of promotional effort at the optimum, the incentive compatibility constraints implies $\alpha > 0$ and $\beta > 0$. In words, the insurers should condition the sales commissions at the same time on collected premiums and on indemnity payments. Because of the risk premium paid to the agent, the expected profit of the insurer is lower than when he observes e and k. The equilibrium levels of e and k also differ from their perfect information levels e^* and k^*. Lastly, at a free entry equilibrium, the number of firms in the market is lower than when the insurer has perfect information about his agent's activity.

Insurance fraud through collusion between policyholders and agents may also occur in the claims settlement phase, particularly in an independent agency system. As emphasized by Mayers and Smith (1981), independent agents usually are given more discretion in claims administration than exclusive agents and they may intercede on the policyholder's behalf with the company's claims adjuster. Influencing claims settlement in the interest of their customers is all the more likely that independent agents may credibly threat to switch their business to another insurer.

Claims fraud at the claims settlement stage may also go through more complex collusion schemes involving policyholders, agents and adjusters. Rejesus *et al.* (2004) have analyzed such collusion patterns in the US Federal Crop Insurance Program. Here the policyholders are farmers and the loss is the difference between the actual yield at harvest and the guaranteed yield specified in the insurance contract. Farmers may collude with agents and adjusters to manipulate the size of the loss in order to increase the indemnity. An agent is paid a percentage of the premiums from all insurance policies he sells. An adjuster is paid on the basis of the number of acres he adjusts. Farmers, agents and adjusters have two possible types : they may be honest or dishonest. Only dishonest individuals may collude. Dishonest agents can potentially have customers from two populations (honest and dishonest farmers), while honest agents only sell policies to honest producers. Thus, the main benefit of collusion to dishonest agents is the chance to have a larger customer pool. Both honest and dishonest adjusters can work for honest and dishonest agents. However a dishonest adjuster can work for a dishonest agent on all of his policyholders (both honest and dishonest). On the contrary, an honest adjuster can only work on the dishonest agent's honest policyholders, but not the dishonest policyholders. Therefore, a dishonest adjuster has a larger customer base. Thus the opportunity to adjust more acres and earn more money is the main benefit of collusion to adjusters.

Rejesus *et al* (2004) consider various patterns of collusion, including "collusion with intermediary", nonrecursive collusion and bilateral collusion. The "cart wheel" model of collusion is an example of collusion with intermedi-

aries. It is based on the principle of linked actions going from a central group of conspirators (the "cart wheel" hub) to many actors (the "rim") through a network of conspiracy intermediaries (the "spokes" in the wheel). Rejesus *et al* (2004) report that, according to compliance investigators of the United States Department of Agriculture (USDA), the structure of collusion in the Federal Crop Insurance Program is configured as such a cart wheel conspiracy, where agents may be the hub, adjusters may be the spokes and farmers may be the rim. Agents, adjusters and farmers may also be linked to one another nonrecursively, contrary to the cart wheel model where there is an intermediary that links the two other actors. Furthermore, collusion may also exist between two individuals rather than three.

The authors use data of the USDA's Risk Management Agency (RMA) to flag anomalous individuals[55]. They show that the pattern of collusion that best fits the data is the nonrecursive scheme. Hence, coordinated behavior between the three entities seems to be the most likely pattern of collusion. The second best pattern of collusion is the collusion with intermediary, where the farmer is the link to both the agent and the adjuster. An example of this type of collusion pattern is the "kickback" scheme, in which a farmer initiates two separate side-contracts with the adjuster and the agent and where he promises them kickbacks from fraudulent claims. The results of Rejesus *et al* (2004) are in contrast to the RMA investigators'belief that the most prevalent pattern of collusion is where the adjuster is the one who initiates and coordinate the collusion as in the above description of the cart wheel pattern.

11 Collusion with service providers

Claims fraud may go through collusion between policyholders and service providers (e.g., car repairers, hospitals, etc). During the two last decades, concentration in the insurance market and in the markets for related services went along with the creation of affiliated service providers networks. This includes managed care organizations for health insurance (such as HMO and PPO in the US) or Direct Repair Programs (DRP) for automobile insurance. Insurance companies may choose to have a restrained set of affiliated service providers for various reasons, including decreasing claims handling

[55]Rejesus *et al* (2004) use indicators of anomalous outcomes. Some of them are applicable to the three types of agents (e.g., the indemnity/ premium ratio), others are specific to agents (e.g., the fraction of policies with loss in the total number of policies sold by the agent) or to adjusters (e.g., the indemnity per claim for the adjuster divided by average adjusted claims in the county).

costs, monitoring providers more efficiently or offering more efficient incentive schemes to providers; see particularly Gal-Or (1997), and Ma and McGuire (1997),(2002) in the case of managed health care.

Bourgeon *et al.* (2008) have analyzed how service providers networks may act as a device to fight claims fraud, when there is a risk of collusion between providers and policyholders[56].They limit attention to a simple setup of a double vertical duopoly with two insurance companies and two service repairers. Providers compete on a horizontally differentiated market modelled as the Hotelling line (providers are not valued the same by policyholders) where they have some market power because of the imperfect substituability of their service. Insurers are perceived as potentially perfectly substitutable by individuals, but they may require their customers to call in a specific provider (say a car repairer) in case of an accident. Two main affiliation structure are considered. In the case of non-exclusive affiliation (Figure 11), customers of both insurance companies are free to choose their providers, while under exclusive affiliation (Figure 12), insurance companies are attached to their own providers[57]. When there is no risk of collusion between providers, exclusive affiliation allows to transfer some market power from the differentiated providers to the undifferentiated insurers, and that transfer will be a disadvantage for the customers. In this case, Bourgeon *et al.* (2008) show that exclusive affiliation is the most likely structure that may emerge in such a setting, with a negative effect on the customers welfare and higher insurers'profit. Hence, if the government gives more social value to the insured's welfare (in terms of wealth certainty equivalent) than to insurers' profit, then it should prevent insurers to restrict access to providers.

Figure 11

Figure 12

Providers and policyholders may collude to file fraudulent claims. We may for instance think of a car repairer who would facilitate fraudulent claiming by certifying that a policyholder actually needed a repair, although that was not the case. Such a collusion may be detered through auditing. Let us assume that providers are risk-neutral. Collusion will be detered if the expected gains obtained by a provider from a collusive deal (i.e. the fraction

[56]See also Brundin and Salanié (1997).

[57]Bourgeon *et al.* (2008) also consider the case of common affiliation in which insurers insurers choose the same provider as their unique referral, and the case of asymmetric affiliation in which one insurer is affiliated with one single provider while customers of the other insurer are free to call in the provider they prefer.

of the insurance indemnity he would receive) is lower than the expected fines he would have to pay if audit reveals collusion. Thus, collusion proofness may lead insurers to reduce their coverage in order to decrease the collusion stake, hence a welfare loss for risk-averse policyholders. Bourgeon *et al.* (2008) show that in a one-shot setting this collusion-proofness condition does not modify the previous conclusion : the defence of the policyholders'interests may still legitimately lead the government to prohibit exclusive affiliation regimes. Matters are different when insurers and providers are engaged in a repeated relationship. In such a setting, a provider is detered from colluding with a customer if his loss in case of an audit is sufficiently large and the threat of retaliation credible. Assume insurers offer insurance contracts that would not be collusion-proof in a one-period framework. Under non-exclusive affiliation, retaliation against a malevolent provider is possible only if insurers agree to punish him simultaneously in the future periods, say by excluding him from their networks or by switching to collusion-proof insurance contracts for all policyholders who would choose this provider[58]. This would require a high degree of coordination between insurers. The situation is different under exclusive affiliation. In particular, if an insurer comes back to collusion-proof contracts after a fraud has been detected (while its competitor does not modifies its offer), then its provider's future profit is reduced. When providers put sufficiently large a weight on future profits, i.e., when their discount factor is large enough, this threat destroys the incentives to collude, even if the probability of detecting collusion is low or when the fines imposed on revealed defrauders are low. In other words, exclusive affiliation may complement imperfect auditing. It may also supplement an inefficient judicial system, where defrauders can easily avoid being strongly fined because insurers have difficulty providing strong evidence in court.

Note finally that detering collusion between policyholders and service providers may not be optimal if some providers are collusive while some are honest. Indeed, if insurers cannot distinguish collusive providers from honest ones, they must either separate them through self-selection contracts or offer collusion-proof contracts to all providers. Both solutions involve distortions in ressource allocation. Alger and Ma (2003) consider such a model, with two types of providers. If the insurer is unable to screen providers by offering them a menu of self-selection contracts, then collusion is tolerated if and only if the provider is collusive with a sufficiently low probability.[59]

[58] Indeed, under non-exclusive affiliation, if there is only one insurance company (the one that has detected collusion) that excludes the defrauder from its network or that switches to collusion-proof contracts, then insureds will move to its competitor and the malevolent provider will not be affected.

[59] Alger and Ma (2003) do not obtain the same result when the insurer can use menus

12 Conclusion

Although the theory of insurance fraud is far from being complete, this survey allows us to draw some tentative conclusions. Firstly, insurance fraud affects the design of optimal insurance policies in several ways. On the one hand, because of claims' monitoring costs, an optimal contract exhibits non-verification with constant net payouts to insureds in the lower loss states and (possibly random) verification for some severe losses. In some cases, a straight deductible contract is optimal. On the other hand, the possibility for policyholders either to manipulate audit costs or to falsify claims should lead insurers to offer contracts that exhibit some degree of coinsurance at the margin. The precise form of coinsurance depends on the specification of the model. For instance, it may go through a ceiling on coverage or through over-compensation for small losses and undercompensation for large losses. However, the fact that insurers should not be offered policies with full insurance at the margin seems a fairly robust result as soon as they may engage in costly activities that affect the insurer's information about damages. Secondly, insurance fraud calls for some cooperation among insurance companies. This may go through the development of common agencies that build data bases about past suspicious claims, that develop quantitative method for better detecting fraudulent claims[60] and that spread information among insurers. In particular data bases may help to mitigate the inefficiency associated with adverse selection, that is with the fact that insurers are unable to distinguish potential defrauders from honest policyholders. Cooperation among insurers may also reduce the intensity of the credibility constraints that affect antifraud policies. Free-riding in antifraud policies could be analyzed along the same lines and it also calls for more cooperation among insurers. Thirdly, insurance fraud frequently goes through collusion with a third party, be it an insurance agent or a service provider. Contractual relationships between insurers and these third parties strongly affects the propensity of policyholders to engage in insurance fraud activities. In particular, conditioning sales commissions paid to agents on a loss-premium ratio results from a compromise between two objectives: providing incentives to make promotional effort and detering collusion with customers. Risk premiums borne by agents are then an additional cost of the distribution system, which ultimately affects the efficiency of insurance industry. Preventing collusion between a policyholder and his own agent is a still more difficult challenge. Vertical integration of these agents by insurance companies (for instance through affiliated auto-

of contracts.

[60] See Derrig and Ostaszewski (1995), Artis *et al.* (1999), Viaene *et al.* (2002).

mechanics networks) is likely to mitigate the intensity of collusion in such cases.

Appendix

Proof of Lemma 1

Let

$$\begin{aligned}
&\widetilde{t}(x) = Sup\{t(x), t(y), y \in M^c\},\\
&t_0 = Inf\{\widetilde{t}(x), x \in [0, \overline{x}]\},\\
&\widetilde{M} = \{x \mid \widetilde{t}(x) > t_0\},\\
&\widetilde{P} = P.
\end{aligned}$$

Obviously, the contract $\widetilde{\delta} = \{\widetilde{t}(.), \widetilde{M}, \widetilde{P}\}$ is incentive compatible. Hence $\widetilde{\delta}$ and δ yield the same insurance payment.

Let $\widehat{x}(x)$ be an optimal claim of the policyholder under δ when he suffers a loss x. Let $x_0 \in \widetilde{M}$. We then have $\widetilde{t}(x_0) > \widetilde{t}(x_1)$ for some x_1 in $[0, \widetilde{x}]$. This gives $\widehat{x}(x_0) \in M$, otherwise $\widehat{x}(x_0)$ would be a better claim than $\widehat{x}(x_1)$ under δ when $x = x_1$. Audit costs are thus lower under $\widetilde{\delta}$ than under δ.

Proof of Lemma 2 [61]

Let

$$\mathcal{L} = U(W - P - x + t(x))f(x) + \lambda[t(x) + c] \text{ if } x \in M$$

be the Lagrangean, with λ a multiplier associated with the non-negative expected profit constraint. When P, t_0 and M are fixed optimally, the schedule $t(.) : M \to R_+$ is such that

$$\frac{\partial \mathcal{L}}{\partial t} = U'(W - P - x + t(x))f(x) - \lambda f(x) = 0.$$

This allows us to write

$$t(x) = x - k \text{ for all } x \in M,$$

where k is a constant.

Assume there exist $0 \leq a_1 < a_2 < a_3 < a_4 \leq \overline{x}$ such that

$$\begin{aligned}
&[a_1, a_2) \cup (a_3, a_4] \subset M,\\
&(a_2, a_3) \subset M^c.
\end{aligned}$$

[61] This proof follows Bond and Crocker (1997).

Let

$$\begin{aligned} M_* &= M - \{[a_1, a_2) \cup (a_3, a_4]\} \\ M_*^c &= M^c - [a_2, a_3] \end{aligned}$$

We have

$$\begin{aligned} EU &= \int_{M_*} U(W - P - k)dF(x) + \int_{M_*} U(W - P - k + t_0)dF(x) \\ &+ \int_{a_1}^{a_2} U(W - P - k)dF(x) + \int_{a_2}^{a_3} U(W - P - k + t_0)dF(x) \\ &+ \int_{a_3}^{a_4} U(W - P - k)dF(x) \end{aligned} \tag{41}$$

and

$$\begin{aligned} E\Pi = P &- \int_{M_*} (x - k + c)dF(x) - \int_{M_*^c} t_0 dF(x) \\ &- \int_{a_1}^{a_2} (x - k + c)dF(x) - \int_{a_2}^{a_3} t_0 dF(x) \\ &- \int_{a_2}^{a_4} (x - k + c)dF(x) = 0. \end{aligned} \tag{42}$$

Differentiating (42) with respect to a_2 and a_4 gives

$$da_3 = \frac{(a_2 - k + c - t_0)f(a_2)da_2}{a_3 - k + c - t_0}$$

which implies

$$dEU = f(a_2)\Delta(t_0 - a_2 + k - c)da_2$$

with

$$\Delta = \frac{U(W - k - P) - U(W - P - a_3 + t_0)}{a_3 - k - t_0 + c} - \frac{U(W - k - P) - U(W - P - a_2 + t_0)}{a_2 - k - t_0 - c}$$

The concavity of U guarantees that $\Delta > 0$. Furthermore $a_2 - k \geq t_0$ since $[a_1, a_2) \subset M$. We thus have $dEU > 0$ if $da_2 < 0$.

Proof of Proposition 1

Let us delete the constraint (6). We may check that it is satisfied by the optimal solution of this less constrained problem. Assigning a multiplier $\lambda \geq 0$

to the non-negative profit constraint, the first-order optimality conditions on k, P and m are respectively

$$[1 - F(m)][U'(W - P - k) - \lambda] = 0 \tag{43}$$

$$\int_0^m U'(W - x - P)dF(x) + [1 - F(m)]U'(W - P - k) = \lambda \tag{44}$$

$$\begin{aligned} & U(W - m - P)f(m_+) - U(W - P - k)f(m_+) + \lambda(c + m - k)f(m_+) \\ \leq\ & 0 \\ =\ & 0 \text{ if } m > 0 \end{aligned} \tag{45}$$

(43), (44) and $F(m) \geq f(0) > 0$ for all $m \geq 0$ give

$$U'(W - P - k) = \frac{1}{F(m)} \int_0^m U'(W - x - P)dF(x)$$

which implies $0 < k < m$ if $m > 0$ and $k = 0$ if $m = 0$.

Assume $m = 0$. Substituting $k = m = 0$ in (45) then gives $\lambda c f(0_+) \leq 0$, hence a contradiction.

Proof of Proposition 2

The first-order optimality conditions on k, P and t_0 are respectively

$$[1 - F(m)][U'(W - P - k) - \lambda] \tag{46}$$

$$f(0)U'(W-P) + \int_{0_+}^m U'(W-x-P+t_0)dF(x) + [1-F(m)]U'(W-P-k) = \lambda \tag{47}$$

$$\int_{0_+}^m U'(W - x - P + t_0)dF(x) = \lambda[F(m) - f(0)] \tag{48}$$

(46), (47), (48) and $F(m) \geq f(0) > 0$ for all $m \geq 0$ give $k = 0$ and $\lambda = U'(W - P)$. Using (48) then yields

$$[F(m) - f(0)]U'(W - P) = \int_{0_+}^m U(W - x - P + t_0)dF(x)$$

which implies $0 < t_0 < m$ if $m > 0$.

Consider m as a fixed parameter. Let $\Phi(m)$ be the optimal expected utility as a function of m. The envelope theorem gives

$$\begin{aligned} & \Phi'(m) = U'(W - m - P + t_0)f(m) - U(W - P - k)f(m) \\ & +\lambda(t_0 + c + m - k)f(m) \end{aligned}$$

if $m > 0$. When $m \to 0$, then $t_0 \to 0$. Using $k = 0$ then gives

$$lim_{m \to 0} \Phi'(m) = \lambda c f(0_+) > 0$$

which implies $m > 0$ at the optimum.

Proofs of Proposition 3 and 5. See Picard (2000).

Proof of Proposition 4 and 6. See Bond and Crocker (1997).

Proof of Proposition 7. See Mookherjee and Png (1989) and Fagart and Picard (1999).

Proof of Proposition 8. See Fagart and Picard (1999).

Proof of Proposition 9 to 12. See Picard (1996)

Proof of Proposition 13

Optimality conditions are written as

$$\begin{aligned} \widehat{p}_i &= 1 \text{ if } cq_i^n - \lambda q_i^f < 0, \\ \widehat{p}_i &\in [0,1] \text{ if } cq_i^n - \lambda q_i^f = 0, \\ \widehat{p}_i &= 0 \text{ if } cq_i^n - \lambda q_i^f > 0, \end{aligned}$$

where λ is a Lagrange multiplier. i^* is the smallest index i in $\{1, ..., \ell\}$ such that $q_i^f / q_i^n \geq c/\lambda$.

References

Alger, I. and C.A. Ma (2003). "Moral hazard, insurance and some collusion", *Journal of Rconomic Behavior and Organization*, 50, 225-247.

Andreoni J., B. Erard and J. Feinstein (1998). "Tax compliance", *Journal of Economic Literature*, XXXVI, 818–860.

Arrow, K. (1971). *Essays in the Theory of Risk Bearing*, North-Holland, Amsterdam.

Artis, M., M. Ayuso and M. Guillen (1999). "Modelling different types of automobile insurance fraud behaviour in the Spanish market", *Insurance : Mathematics and Economics*, 24, 67-81.

Baron, D. and D. Besanko (1984). "Regulation, asymmetric information and auditing", *Rand Journal of Economics*,15 (4), 447–470.

Becker, G. (1968). "Crime and punishment: an economic approach", *Journal of Political Economy*,76,169–217.

Bond, E. and K.J. Crocker (1997). "Hardball and the soft touch: the economics of optimal insurance contracts with costly state verification and endogenous monitoring costs", *Journal of Public Economics*, 63, 239–264.

Bourgeon, J.M. and P. Picard (1999). "Reinstatement or insurance payment in corporate fire insurance", *Journal of Risk and Insurance*, 67(4), 507-526.

Bourgeon, J.M. and P. Picard (2012). "Fraudulent claims and nitpicky insurers", Working Paper N° 2012-06, Ecole Polytechnique, Department of Economics.

Bourgeon, J.M., P. Picard and J. Pouyet (2008), "Providers' affiliation, insurance and collusion", *Journal of Banking and Finance*, 32, 170-186.

Boyer, M. (1999). "When is the proportion of criminal elements irrelevant? A study of insurance fraud when insurers cannot commit", in *Automobile Insurance : Road Safety, New Drivers, Risks, Insurance Fraud and regulation*, Edited by G. Dionne and C. Laberge-Nadeau, Kluwer Academic Publishers.

Boyer, M. (2000a). "Insurance taxation and insurance fraud", *Journal of Public Economic Theory*, 2(1), 101-134.

Boyer, M. (2000b). "Centralizing insurance fraud investigation", *Geneva Papers on Risk and Insurance Theory*, 25, 159-178.

Boyer, M. (2001). "Mitigating insurance fraud : lump-sum awards, premium subsidies, and indemnity taxes", *Journal of Risk and Insurance*, 68(3), 403-436.

Boyer, M. (2004). "Overcompensation as a partial solution to commitment and renegotiation problems : the case of *ex post* moral hazard", *Journal of Risk and Insurance*, 71(4), 559-582.

Brundin, I. and Salanié, F. (1997). "Fraud in the insurance industry: an organizational approach", mimeo, Université de Toulouse.

Clarke, M. (1997). *The Law of Insurance Contracts*, Lloyd's of London Press Ltd.

Crocker, K.J. and J. Morgan (1997), "Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts", *Journal of Political Economy*, 106(2), 355–375.

Crocker, K. J. and S. Tennyson (1999). "Costly state falsification or verification? Theory and evidence from bodily injury liability claims", in *Automobile Insurance : Road Safety, New Drivers, Risks, Insurance Fraud and regulation*, Edited by G. Dionne and C. Laberge-Nadeau, Kluwer Academic Publishers.

Crocker, K.J. and S. Tennyson (2002), "Insurance fraud and optimal claims settlement strategies", *Journal of Law and Economics*, vol. XLV, 469-507.

Cummins, J.D. and S. Tennyson (1992). "Controlling automobilee insurance costs", *Journal of Economic Perspectives*, 6, N°2,95-115.

Cummins, J. D. and S. Tennyson (1994). "The tort system 'lottery' and insurance fraud: theory and evidence from automobile insurance", *mimeo*, The Wharton School, University of Pennsylvania, 94-05.

Darby, M. and E. Karni (1973). "Free competition and the optimal amount of fraud", *Journal of Law and Economics*, 16, 67–88.

Dean, D.H. (2004). "Perceptions olf the ethicality of consumers'attitudes toward insurance fraud", *Journal of Business Ethics*, 54, 1, 67-79.

Derrig, R. and K. Ostaszewski (1995). "Fuzzy techniques of pattern recognition in risk and claim classification", *Journal of Risk and Insurance*, 62, 447–482.

Derrig, R.A., W. H. and X. Chen (1994). "Behavioral factors and lotteries under no-fault with a monetary threshold: a study of massachusetts automobile claims", *Journal of Risk and Insurance*, 9(2), 245–275.

Dionne, G. (1984). "The effects of insurance on the possibilities of fraud", *The Geneva Papers on Risk and Insurance*, 9(32), 304–321.

Dionne, G. and R. Gagné (2001). "Deductible contracts against fraudulent claims : evidence from automobile insurance", *Review of Economics and Statistics*, 83, 2, 290-301.

Dionne, G. and R. Gagné (2002). "Replacement cost endorsement and opportunistic fraud in automobile insurance", *Journal of Risk and Uncertainty*, 24, 213-230.

Dionne, G., F. Giuliano and P. Picard (2009), "Optimal auditing with scoring : theory and application to insurance fraud", *Management Science*, 55, 58-70.

Dionne, G. and P. St-Michel (1991)."Workers' compensation and moral hazard", *Review of Economics and Statistics*, 83(2), 236–244.

Dionne, G., P. St-Michel and C. Vanasse (1995). "Moral hazard, optimal auditing and workers' compensation" in *Research in Canadian Workers' Compensation*, T. Thomason and R.P. Chaykowski eds, IRC Press, Queen's University at Kingston, 85-105.

Dionne, G. and K.C. Wang (2011). "Does opportunistic fraud in automobile theft insurance fluctuate with the business cycle ?", *mimeo*, CIRRELT, Working Paper N° 2011-49.

Dixit, A. (2000). "Adverse selection and insurance with *Uberrima Fides*", in *Incentives, Organization and Public Economics : Essays in Honor of Sir*

James Mirrlees, Edited by P.J. Hammond and G.D. Myles, Oxford University Press, Oxford.

Dixit, A. and P. Picard (2003). "On the role of good faith in insurance contracting", in Economics for an Imperfect World, Essays in Honor of Joseph Stiglitz, Edited by R. Arnott, B. Greenwald, R. Kanbur et B. Nalebuff, MIT Press, Cambridge, MA, 17-34.

Fagart, M. and P. Picard (1999). "Optimal insurance under random auditing", *Geneva Papers on Risk and Insurance Theory*, 29, 1, 29–54.

Fudenberg, D., D.M. Kreps and E. Maskin (1990). "Repeated games with long-run and short-run players", *Review of Economic Studies*, 57(4), 555-573.

Fukukawa, K., Ennew,C. and S. Diacon (2007). "An eye for an eye : investigating the impact of consumer perception of corporate unfairness on aberrant consumer behavior", in *Insurance Ethics for a more Ethical World (Research in Ethical Issues in Organizations)*, Edited by P. Flanagan, P. Primeaux and W. Ferguson, Vol. 7, 187-221, Emerald Group Publishing Ltd.

Gal-Or, E. (1997). "Exclusionary equilibria in healthcare markets", *Journal of Economics and Management Strategy*, 6: 5-43.

Gollier, C. (1987). "Pareto-optimal risk sharing with fixed cots per claim", *Scandinavian Actuarial Journal,* 62–73.

Graetz, M.J., J.F. Reinganum and L.L. Wilde (1986). "The tax compliance game: toward an interactive theory of law inforcement", *Journal of Law, Economics and Organization,* 2(1), 1–32.

Guesnerie, R. and J.J. Laffont (1984). "A complete solution to a class of principal-agent problems, with an application to the control of a self-managed firm", *Journal of Public Economics,* 25, 329–369.

Hau, A. (2008). "Optimal insurance under costly falsification and costly inexact verification", *Journal of Economic Dynamics and Control*, 32, 5, 1680-1700.

Holmström, B. (1979). "Moral hazard and observability", *Bell Journal of Economics*, 10(1), 79–91.

Huberman, G., D. Mayers and C.W. Smith Jr. (1983). "Optimum insurance policy indemnity schedules", *Bell Journal of Economics,* 14, Autumn, 415–426.

Kim, W.-J., D. Mayers and C.W. Smith, Jr.(1996). "On the choice of insurance distribution systems" *Journal of Risk and Insurance"*, 63(2), 207–227.

Krawczyk, M. (2009). "The role of repetition and observability in deterring insurance fraud", *Geneva Risk and Insurance Review*, 34, 74-87.

Lacker, J. and J.A. Weinberg (1989). "Optimal contracts under costly state falsification", *Journal of Political Economy*, 97, 1347–1363.

Ma, C. A., and T. McGuire (1997). "Optimal Health Insurance and Provider Payment", *The American Economic Review*, 87(4): 685-704.

Ma, C. A., and T. McGuire (2002). "Network Incentives in Managed Health Care", *Journal of Economics & Management Strategy*, 11(1): 1-35.

Maggi, G. and A. Rodriguez-Clare (1995). "Costly distortion of information in agency problems" Rand Journal of Economics", 26, 675–689.

Mayers, D. and C.S. Smith, Jr. (1981). "Contractual provisions, organizational structure, and conflict control in insurance markets", *Journal of Business,* 54, 407–434.

Melumad, N. and D. Mookherjee (1989). "Delegation as commitment: The case of income tax audits", *Rand Journal of Economics,* 20(2), 139–163.

Miyazaki, A.D. (2009). "Perceived ethicality of insurance claim fraud : do higher deductibles lead to lower ethical standards ?", *Journal of Business Ethics*, 87, 4, 589-598.

Mookherjee, D. and I. Png (1989). "Optimal auditing insurance and redistribution", *Quarterly Journal of Economics*, CIV., 205–228.

Moreno, I., F.J. Vasquez and R. Watt (2006). "Can bonus-malus allievatez insurance fraud ?", *Journal of Risk and Insurance*, 73, 1, 123-151.

Myerson, R. (1979). "Incentive compatibility and the bargaining problem", *Econometrica,* 47, 61–74.

Picard, P. (1996). "Auditing claims in insurance market with fraud: the credibility issue", *Journal of Public Economics*, 63, 27–56.

Picard, P. (2000). "On the design of optimal insurance contracts under manipulation of audit cost", *International Economic Review*, 41, 1049-1071.

Picard, P. (2009). "Costly risk verification without commitment in competitive insurance markets", *Games and Economic Behavior*, 66, 893-919.

Puelz, R. and A. Snow (1997). "Optimal incentive contracting with *ex-ante* and *ex-post* moral hazards: theory and evidence", 14(2),168-188.

Rejesus, R.M., B.B. Little, A.C. Lowell, M. Cross and M. Shucking (2004). "Patterns of collusion in the US Crop Insurance Program : an empirical analysis", *Journal of Agricultural and Applied Economics*, 36, 2, 449-465.

Schiller, J. (2006). "The impact of insurance fraud detection systems", *Journal of Risk and Insurance*, 73,3, 421-438.

Stigler, G. (1970). "The optimal enforcement of laws", *Journal of Political Economy*, 78, 526–536.

Strutton, D., S.J. Vitelle and L.E. Pelton (1994). "How consumers may justify inappropriate behavior in market settings : an application on the techniques of neutralization", Journal of Business Research, 30(3), 253-260.

Tennyson, S. (1997). "Economic institutions and individual ethics: a study of consumer attitudes toward insurance fraud", *Journal of Economic Behavior and Organization*, 32, 247–265.

Tennyson, S. (2002). "Insurance experience and consumers'attitudes toward insurance fraud", *Journal of Insurance Regulation*, 21, 2, 35-55.

Tennyson, S. and P. Salsas-Forn (2002). "Claims auditing in automobile insurance : fraud detection and deterrence objectives", *Journal of Risk and Insurance*, 69, 3, 289-308.

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

Viaene, S. and G. Dedene (2004). "Insurance fraud : issues and challenges", *Geneva Papers on Risk and Insurance*, 29, 2, 313-333.

Viaene,S., R.A. Derrig, B. Baesens and G. Dedene (2002). "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection", *Journal of Risk and Insurance*, 69, 3, 373-421.

Weisberg, H. & Derrig, R. (1993). "Quantitative methods for detecting fraudulent automobile bodily injury claims" ???

Wilson, C. (1977). "A model of insurance markets with incomplete information", *Journal of Economic Theory*, 16, 167–207.

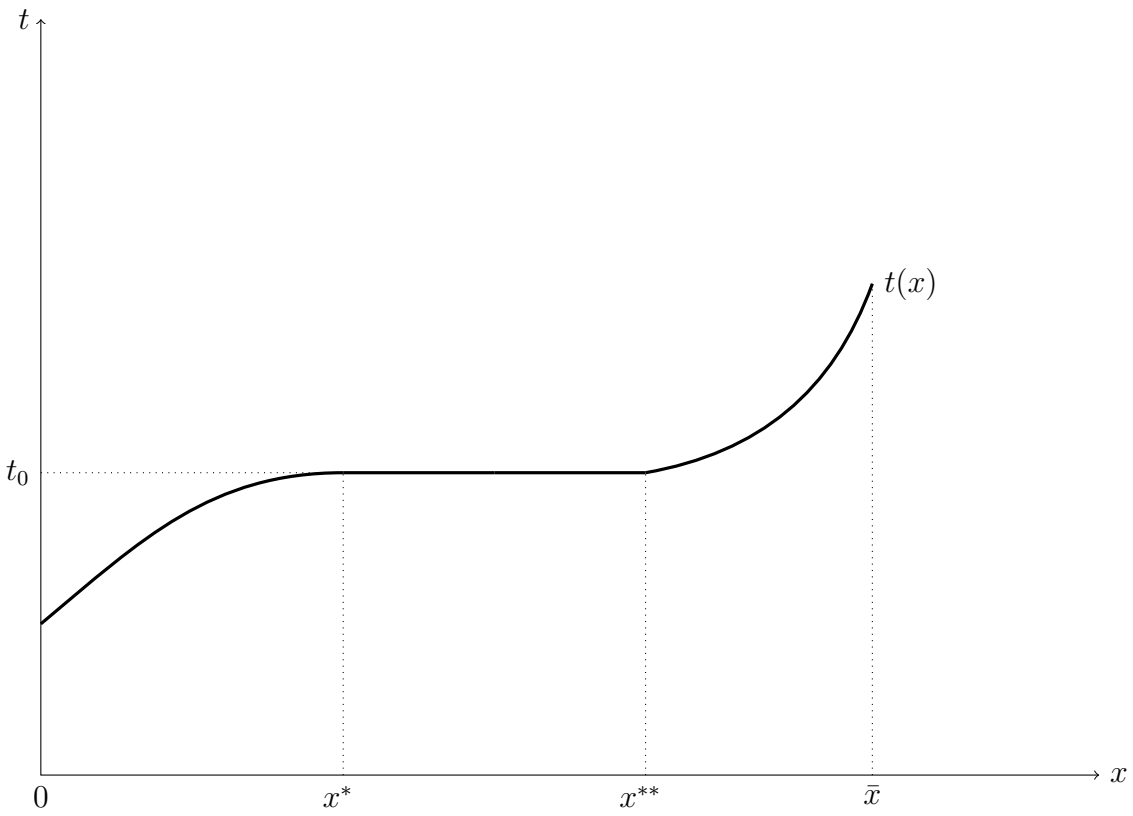

Figure 1: Characterization of incentive compatible contracts

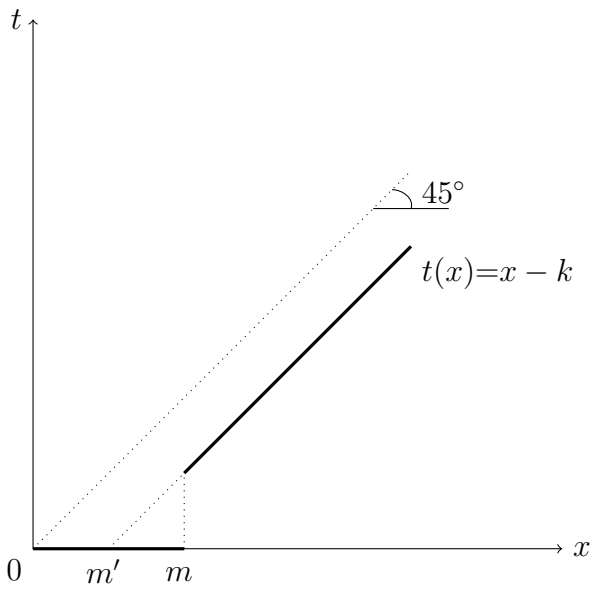

Figure 2: Optimal insurance coverage under deterministic auditing

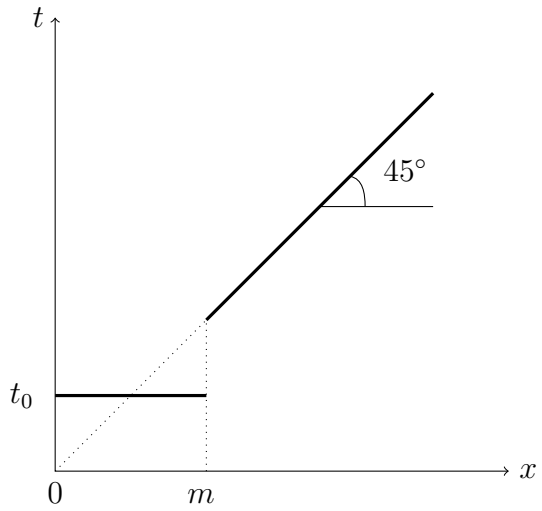

Figure 3: Optimal insurance coverage under deterministic auditing when the insurer can observe whether an accident has occured but not the magnitude of the actual loss

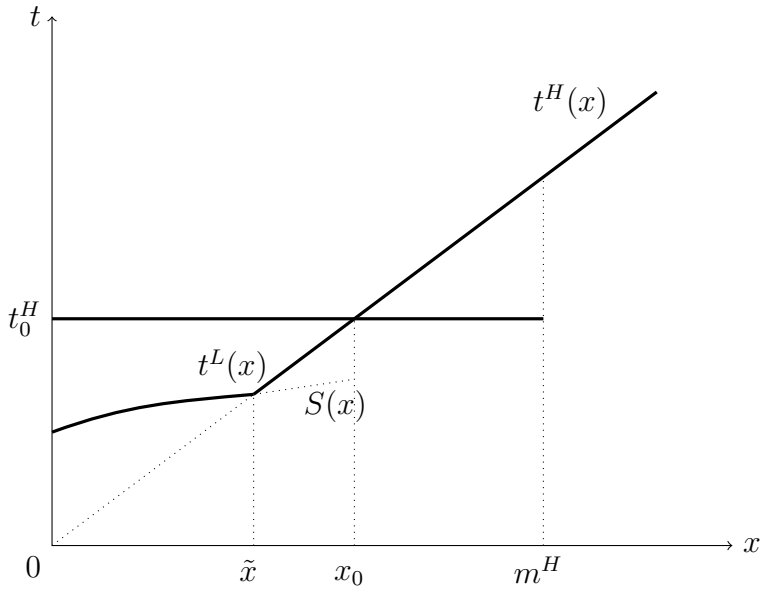

Figure 4: Optimal no-manipulation contract in the Bond-Crocker (1997) model

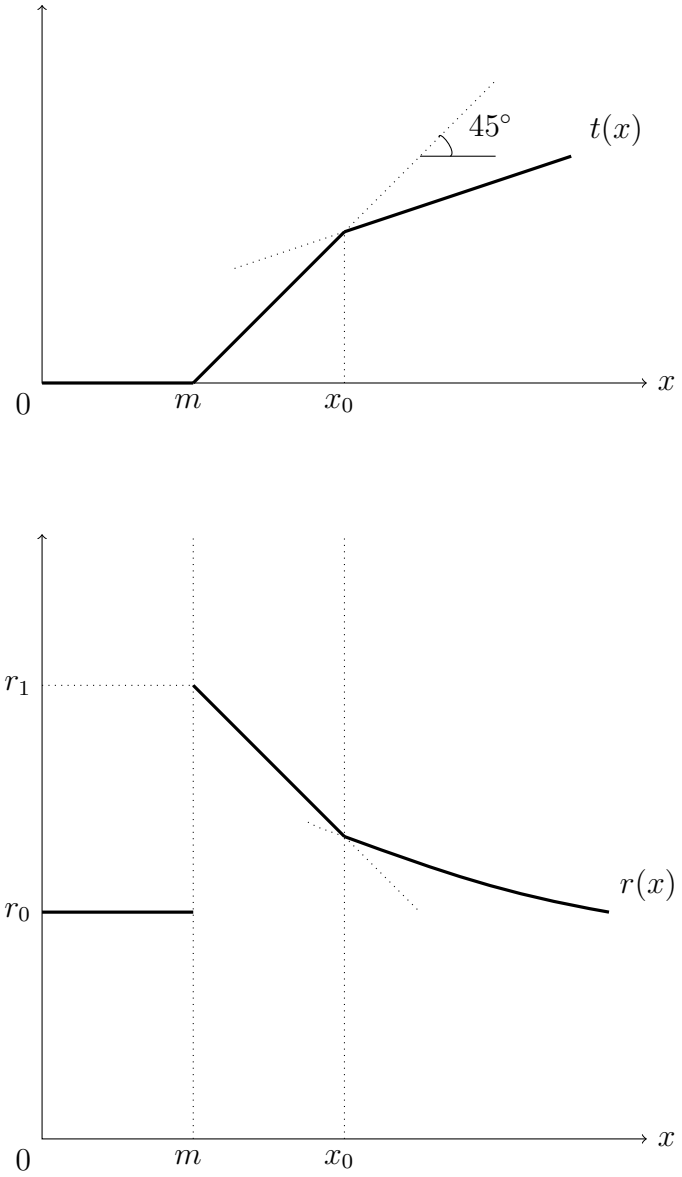

Figure 5: Optimal insurance contract and auditor's contingent fees

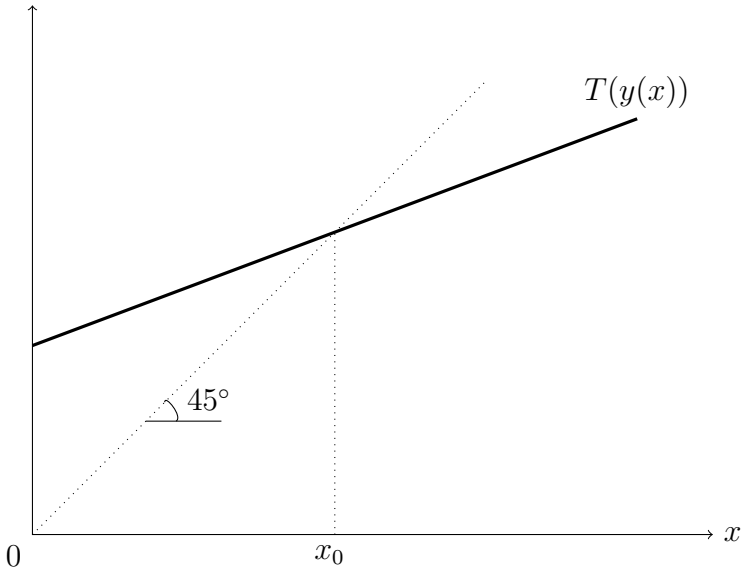

Figure 6: Equilibrium indemnification under costly state falsification

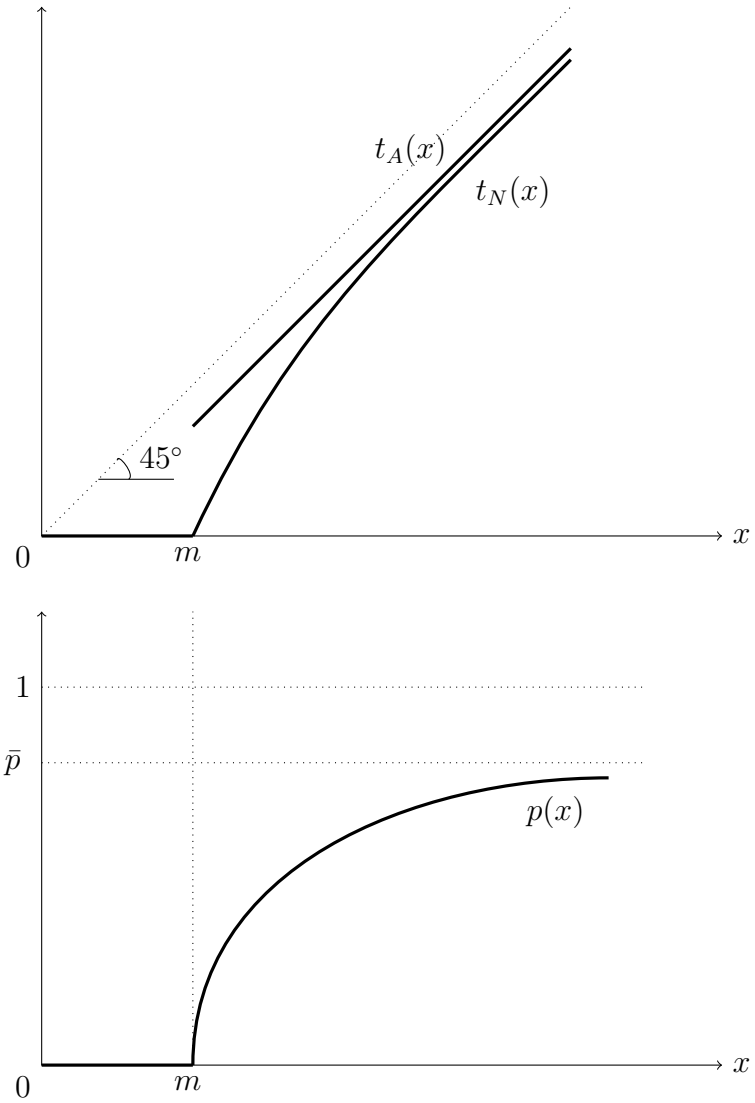

Figure 7: Optimal insurance contract under random auditing when $U(.)$ is CARA

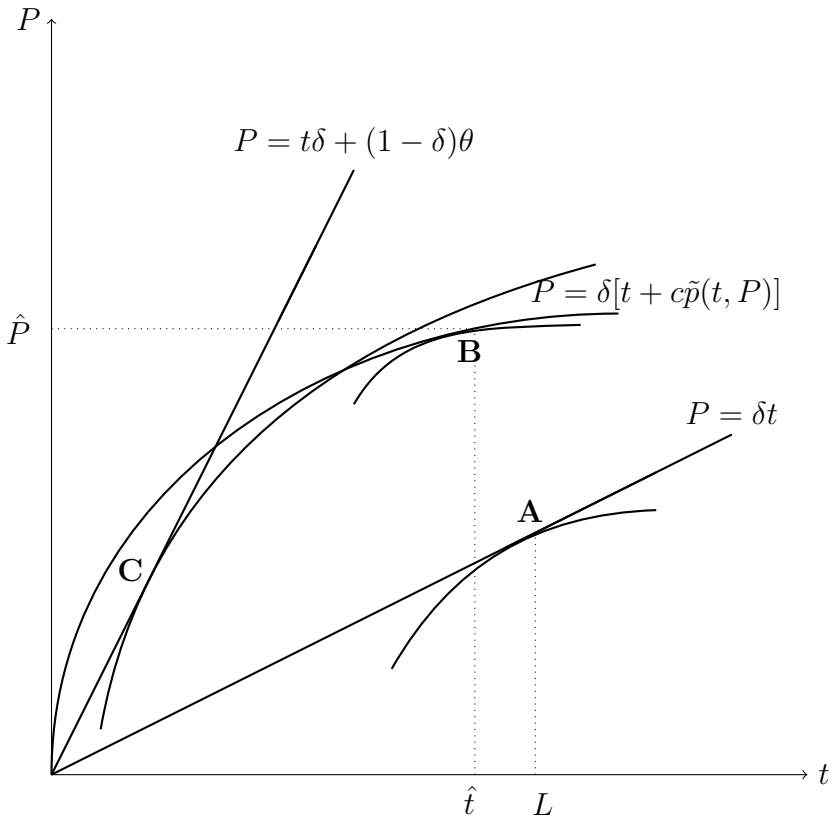

Figure 8: The market equilibrium is at point B when $\theta > \hat{\theta}$

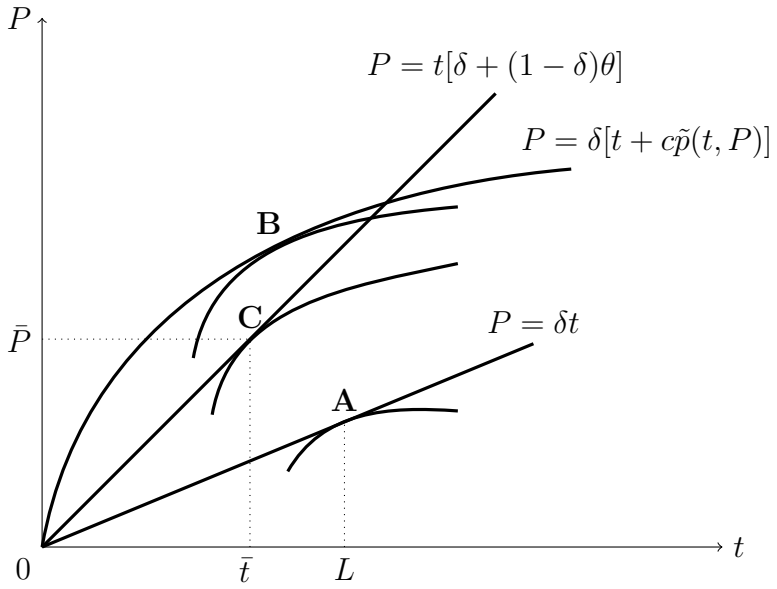

Figure 9: The market equilibrium is at point C when $\theta < \hat{\theta}$

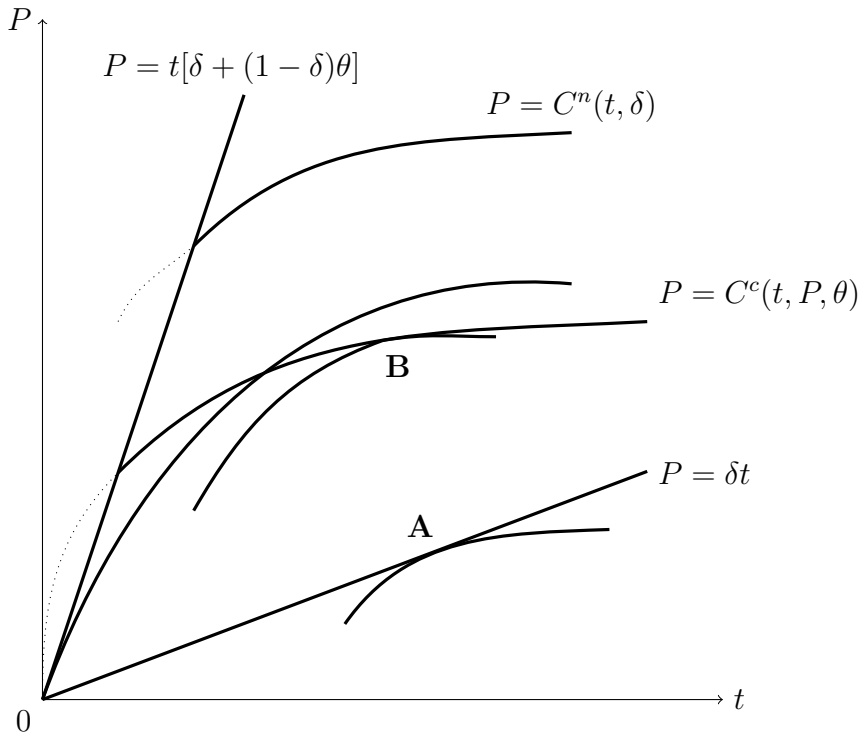

Figure 10: Case where the market shuts down at no-commitment equilibrium

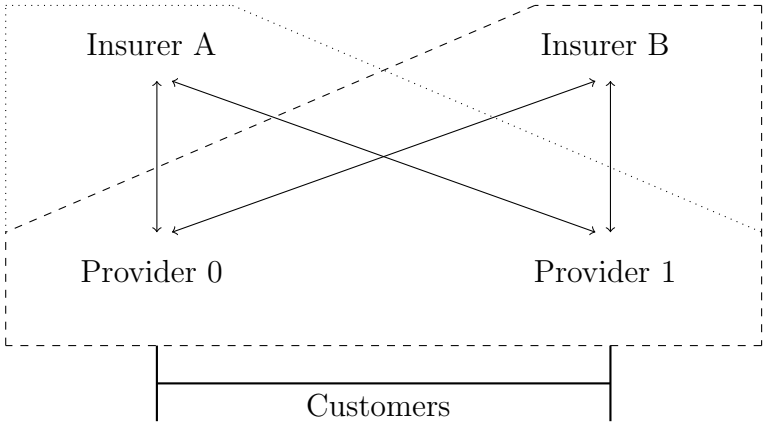

Figure 11: No affiliation

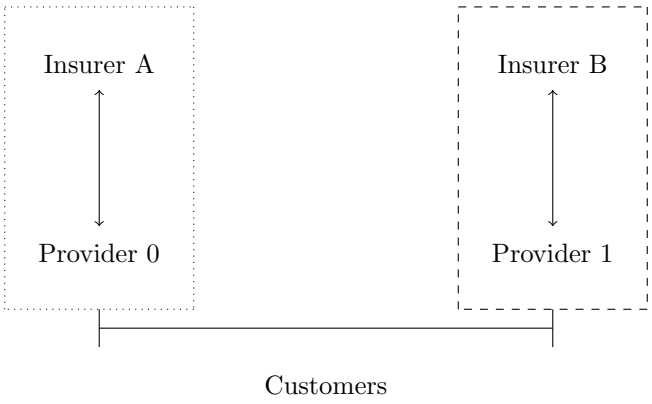

Figure 12: Exclusive affiliation