DeMarzo, Peter M.; Fishman, Michael J.; Hagerty, Kathleen M.

Working Paper

Contracting and enforcement with a self-regulatory organization

CSIO Working Paper, No. 0023

Provided in Cooperation with:
Department of Economics - Center for the Study of Industrial Organization (CSIO), Northwestern University

Suggested Citation: DeMarzo, Peter M.; Fishman, Michael J.; Hagerty, Kathleen M. (2000) : Contracting and enforcement with a self-regulatory organization, CSIO Working Paper, No. 0023, Northwestern University, Center for the Study of Industrial Organization (CSIO), Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/38711

THE CENTER FOR THE STUDY
OF INDUSTRIAL ORGANIZATION
AT NORTHWESTERN UNIVERSITY

Working Paper #0023

Contracting and Enforcement with a Self-Regulatory Organization*

By

Peter M. DeMarzo
Graduate School of Business,
Stanford University

and

Michael J. Fishman and Kathleen M. Hagerty
Kellogg Graduate School of Management,
Northwestern University

October 2000

* We would like to thank Mike Burkart, Gerald Garvey, David Hirshleifer, Chester Spatt, Joel Watson, and other seminar and conference participants for helpful comments and discussions and we would like to thank David Fang for outstanding research assistance.

Abstract

Self regulation is a feature of a number of professions. For example, the government delegates aspects of financial market regulation to self-regulatory organizations (SROs) like the New York Stock Exchange and the National Association of Securities Dealers. We analyze one regulatory task of an SRO, enforcing antifraud rules so agents will not cheat customers. Specifically, we model contracting/enforcement as a two-tier problem. An SRO chooses its enforcement policy: the likelihood an agent is investigated for fraud and a penalty schedule. Given an enforcement policy, agents compete by offering contracts that maximize customers' expected utility. We assume the SRO's objective is to maximize the welfare of its members, the agents. We show that in the static case, the SRO chooses a more lax enforcement policy – less frequent investigations and lower penalties – than what customers would choose. A general conclusion is that control of the enforcement policy governing contracts confers substantial market power to a group of otherwise competitive agents. We also investigate government oversight of the self-regulatory process. The threat of government enforcement leads to more enforcement by the SRO. A similar result applies in a dynamic setting where customers can impose discipline by monitoring an agent's reputation for past performance. Here there are circumstances for which the SRO would choose a more aggressive enforcement policy than is preferred by customers.

1. Introduction

In the U.S., the Securities and Exchange Commission (SEC) has primary responsibility for regulating securities markets. The SEC, however, delegates significant regulatory authority to self-regulatory organizations (SROs), securities industry organizations that are owned and operated by their members. Examples of SROs include the National Association of Security Dealers (NASD), the New York Stock Exchange (NYSE), the Chicago Board Options Exchange, and regional stock and option exchanges. Among the SROs' tasks is to design rules governing their members' practices. In addition, SROs are responsible for enforcing their own rules as well as federal securities laws. They conduct disciplinary proceedings and impose sanctions on members for violations.[1]

Self regulation is also a key feature of other professions, including accounting, law, and (at least up until the late 1980's) medicine (see Shuchman (1981) and Ameringer (1999)). State accounting boards and the American Institute of Certified Public Accountants (AICPA)—these are voluntary membership organizations for accountants—conduct investigations and impose disciplinary actions on accountants who commit fraud or otherwise violate the code of conduct. In some states, state bar associations (also membership organizations) handle attorney disciplinary matters. State medical boards (historically largely controlled by physicians) handle physician disciplinary matters.

Self regulation has always been subject to public criticism. Given that SROs are run for the benefit of their members, there are concerns that SROs have inadequate incentives to enforce rules that protect the public; see for example, Shuchman (1981), McCaffrey and Hart (1998), and Ameringer (1999). How likely is an SRO to investigate one of its members for a violation? What penalties will an SRO set for members who violate the rules? Will the enforcement policy chosen by an SRO coincide with the policy that is preferred by their customers? These are among the questions addressed in this paper.

We consider the enforcement of anti-fraud rules, extending the costly-state-verification model of Townsend (1979) and Mookherjee and Png (1989). In this model, agents facilitate transactions for

[1] See National Association of Securities Dealers (1996), Phillips (1997) and McCaffrey and Hart (1998) for general discussions of SRO enforcement and Frankhauser, et al. (1997) for details of SRO enforcement procedures.

customers but might underreport the payoff from the transaction and keep the remainder. Such fraud can be deterred either by a threat of an investigation and penalty or an incentive contract that pays the agent more when he reports high payoffs. This moral hazard model captures a number of fraudulent activities. For instance, suppose the agent is a broker who executes trades for a customer. The customer does not directly observe the prices at which orders are filled and the broker can cheat by reporting that buy (sell) orders were filled at higher (lower) prices. Alternatively, an agent who manages a customer's trading account may churn the account and collect excessive brokerage fees. Another possibility is that the customer is billed for services not provided, a problem faced by clients of lawyers and doctors. Of course, the agent may simply steal/misuse the customer's money, a problem faced by clients of brokers, accountants, and lawyers.[2]

In this setting, an enforcement policy, consisting of a specification of the likelihood of an investigation and a penalty schedule for fraud, is chosen by the SRO whose objective is to maximize the welfare of its members, the agents. Taking the SRO's enforcement policy as given, agents compete with one another to handle customer transactions. They compete by offering contracts promising (outcome-contingent) payoffs that maximize customers' expected utility. We assume that the SRO cannot control this price competition; indeed, attempts to do so would generally violate antitrust or other laws (e.g., the Securities Acts Amendments of 1975 prohibits exchanges from setting members' commissions).

When choosing an enforcement policy, the SRO anticipates the competition among its members. We show that the SRO mutes this competition by choosing a more lax enforcement policy than is preferred by customers. Investigations for fraud are less frequent and penalties are lower than what a customer would choose (though the revelation principle applies and so in equilibrium there is no fraud). The intuition is that an agent can be induced to be truthful by both the "carrot" of a bonus when he reports a high payoff

[2] McCaffrey and Hart (1998) report that at least 35-40% of NYSE and NASD disciplinary actions and about 12% of all enforcement cases initiated by the SEC involve brokers defrauding customers. Morrison and Wickersham (1998) finds that 9% of physician disciplinary actions in California involved fraud (illegal billing, Medicaid fraud, and theft from patients) and reports that earlier studies find higher percentages of physician fraud among disciplinary actions.

and the "stick" of a threat of investigation and penalty. Agents prefer to be motivated by the carrot rather than the stick. By choosing a lax enforcement policy, an SRO induces customers to offer more of a carrot, i.e., higher bonuses for reporting high payoffs. We also show that a decrease in the cost of an investigation leads an SRO to actually investigate less. Enforcement becomes more aggressive, however, as a customer's alternatives to dealing with an agent of the SRO improve.

A general conclusion of the analysis is that control of the enforcement policy governing contracts confers substantial market power to a group of otherwise competitive agents. In fact, we show that if agents are risk neutral, control of the enforcement policy is equivalent to agents behaving as monopolists.

In standard contracting theory, all contract terms are chosen as part of a bilateral agreement; in particular, contracts specify outcome-contingent payoffs as well as an enforcement policy to govern the contract. In practice, however, parties usually negotiate outcome-contingent payoffs taking the enforcement environment as given. For example, when you hire a broker you agree to the fees to be paid. You do not negotiate the likelihood that the broker is investigated for fraud or the penalty for fraud. Instead you rely on the enforcement policy of some authority, e.g., the NASD, state regulators, and the SEC. The novel aspect of our analysis is to incorporate this feature of contracting. We model contracting/enforcement as a two-tier problem in which an institution determines the enforcement policy taking into account the effect of that policy on the contracts that are subsequently created. Our results highlight how the enforcement institution can affect the division of rents between contracting parties.

We also consider government oversight, meaning that the government observes the SRO's enforcement policy and then chooses its own enforcement policy. We assume the government's objective is to maximize customer expected utility. The threat of government enforcement leads to more aggressive enforcement by the SRO. Moreover, this is achieved without any actual government enforcement.

In the one-period analysis, an agent who is caught defrauding a customer faces a monetary sanction. We also consider a multi-period analysis in which two further penalties become possible. First, an agent who is caught cheating a customer can be barred by the SRO from doing future business. Second, customers can penalize agents by taking their business elsewhere even without direct evidence of fraud.

In effect, the market disciplines agents if customers choose their agent based on past performance. In this dynamic setting, the SRO's optimal enforcement is either aggressive enough to pre-empt this market discipline or is so lax that it relinquishes all discipline to the market. In the former case, the SRO may choose a more aggressive enforcement policy than is preferred by customers.

The self-regulation literature typically deals with two issues. One is whether SROs face enough competition to induce socially efficient self-regulation; see, for example, Pirrong (1995) and Mahoney (1997) who focus on exchanges. Our analysis assumes that while agents of an SRO compete with one another, the SRO itself has some market power. These are natural assumptions for the securities industry and in accounting, law and medicine. The other issue in the literature concerns how SROs exercise market power. For example, Leland (1979), Shaked and Sutton (1981), Saloner (1984) and Gehrig and Jost (1995) examine a self-regulating profession's incentive to limit membership. These papers find that SROs set membership standards that are too high – one could increase efficiency by relaxing standards. In our analysis membership size plays no role. An SRO exercises market power through the way it enforces rules. In contrast with the prior literature, and perhaps more in line with public perception, we show that SRO standards may be too low, with weaker-than-efficient enforcement policies.

Our paper is also related to the hierarchical contracting literature. Faure-Grimaud, Laffont, and Martimort (1999) analyze a two-tier hierarchy; a principal hires an auditor to audit an agent. The principal chooses the auditor's contract and an audit policy and then the auditor offers a contract to the agent. Cremer, Marchand, and Pestieau (1990) and Sanchez and Sobel (1993) analyze hierarchical contracting in tax compliance: the government chooses a tax schedule and then the tax collector chooses an audit policy. In these analyses, like ours, an agent faces a contract and audit policy that are offered by distinct parties (with distinct objectives). In Tirole (1986) and Kofman and Lawarree (1993), a principal chooses contracts for an auditor and agent, and then the auditor and agent can negotiate a side contract motivated by the possibility of issuing a false audit. In these two analyses, the principal anticipates the side contracting opportunities when choosing contracts. In a similar spirit, in our analysis, the SRO chooses an enforcement policy anticipating the contract between the agent and the customer.

The model is described in Section 2. Section 3 analyzes an SRO's optimal enforcement policy and the resulting contract between an agent and customer. Section 4 examines government oversight of an SRO. Section 5 explores the role of reputation in a dynamic setting. Section 6 contains concluding remarks. Proofs of the Propositions appear in the Appendix.

2. The Model

There are a variety of ways to model a customer-agent relation in which the agent might cheat the customer. We use the costly-state-verification model of Townsend (1979) and Mookherjee and Png (1989). An agent observes private information and may misreport this information in order to cheat his customer. For a cost, it is possible to verify the agent's report, and detect the occurrence of fraud.

There are three players: the customer, the agent, and the self-regulatory organization (SRO). The customer hires the agent. The cash flow from using the agent's services is a random variable W that has countable support $\Omega \subset \Re_+$. Assume that Ω has a minimum element, denoted $\underline{w}$. The agent privately observes the realization of the cash flow W and reports it to the customer. This is the source of the moral hazard problem. The agent might lie about the cash flow and keep some of it for himself. The agent can be investigated at a cost of $c \geq 0$. An investigation reveals the realization w of W.

The customer is risk neutral though the agent may be risk averse. This eliminates consideration of a contract that provides the customer with insurance. So in the absence of the moral hazard problem, the optimal contract involves the customer hiring the agent for a fixed fee.[3] The customer's best alternative to transacting with the agent is represented by an expected payoff of α.

Let $u(y)$ denote the agent's utility as a function of income y, where u is increasing, concave, and $u(0) = 0$. The agent has zero initial wealth and faces a limited liability constraint; his net income from the transaction with the customer cannot be less than zero. These assumptions have two implications. First, it

[3] If the customer were risk averse, then since there is no customer moral hazard problem he could share risk with others. With perfect risk sharing with outsiders, the problem is the same as here. On the other hand, since there is an agent moral hazard problem, risk sharing for him is problematic. Thus, agent risk aversion is relevant.

bounds the size of the penalty that can be imposed on the agent. Second, though the agent behaves competitively, he cannot compete away all rents by paying a customer to do business with him.

A contract between the agent and customer is represented by the function z, where $z(r)$ specifies the payment to be made by the agent to the customer if the agent reports that $W = r$. The SRO's enforcement policy specifies an investigation strategy and a penalty schedule. The investigation strategy is represented by the function p, where $p(r)$ specifies the probability that the agent is investigated given a report of r. If an investigation takes place, monetary penalties are assessed according to a penalty schedule x, where $x(w,r)$ specifies the penalty to be paid by the agent given the report r and an actual payoff of w.[4]

In the standard contracting literature, the customer and the agent directly negotiate the contract and enforcement policy (z,p,x). We instead consider an environment in which the enforcement policy (p,x) is first set by the SRO, and then the customer and the agent negotiate the contract z.

The SRO conducts investigations according to the strategy p, incurring cost c whenever an investigation takes place.[5] The SRO collects penalties according to the schedule x from the agent. We restrict the SRO to set penalties $x \geq 0$; that is, we do not allow the SRO to pay the agent. Finally, the SRO charges a transaction fee t that is paid by the customer and used to finance expected enforcement costs net of penalties. (Our results are unchanged if instead the customer collects all or some of the penalties. The SRO would simply adjust the transaction fee. See also footnote 20.)

[4] The enforcement policy depends on the report, r, and not the payment z. In a standard mechanism design approach allowing the policy to depend on z is superfluous. This is not true here since the enforcement policy and contract are chosen by different parties. If the SRO could condition its enforcement policy on z, it could directly control the customer's choice of z by committing to do no enforcement unless z was in some prescribed set. We made this modeling choice because SROs directly control enforcement but do not directly control contracts between members and customers. If an SRO tried to directly control contracts with customers, it would be an antitrust violation.

[5] We discuss p as the probability of an investigation. An alternative interpretation involves imperfect investigations. Suppose investigations are initiated by customer complaints and suppose customers complain unless they receive the best outcome. Then p can be interpreted as the probability that an investigation reveals w, with a cost per investigation of pc (more accurate investigations have higher cost). The SRO chooses the accuracy, p, to employ.

Given an enforcement policy, the customer and agent negotiate the contract, z. Assume the agent behaves competitively (effectively, the customer makes a take-it-or-leave-it offer).[6] The customer and agent each maximize their own expected utility. Anticipating the behavior of the customer and agent, the SRO chooses the enforcement policy (p,x) and transaction fee t to maximize the agent's expected utility.

To summarize, the timing of the problem is as follows:

1. The SRO chooses an enforcement policy (p,x) and transaction fee t.
2. Taking (p,x,t) as given, the customer offers a contract z to the agent.
3. If the agent rejects the contract, the agent receives 0 and the customer receives α. If the agent accepts the contract, the customer pays t and the problem continues.
4. A cash flow w is realized. Only the agent observes this realization.
5. The agent chooses a cash flow, r, to report and gives the customer the corresponding payoff, $z(r)$.[7]
6. Given the report r, the SRO investigates the agent with probability $p(r)$. If an investigation occurs, the SRO pays c and collects the penalty $x(w,r)$ from the agent (subject to the agent's resource constraint).

We now analyze the subgame perfect equilibrium of this game.

2.1 The Agent's Problem

Given (z,p,x), for each outcome w of W the agent chooses a report, $r(w)$, and pays the customer $z(r(w))$. Since the agent has only w available, the reporting strategy, r, is feasible if and only if it satisfies

$$\text{(AF)} \qquad z(r(w)) \leq w \text{ for all } w.$$

Given outcome w and report r, an investigation occurs with probability $p(r)$ and the penalty is $x(w,r)$. Thus, the agent's expected utility is given by

[6] This is without loss of generality since we can trace out the Pareto frontier by varying the reservation payoff α.

[7] If the agent agrees to $z(\cdot)$, then given report r, a payment of $z(r)$ is enforced. Implicitly, we are assuming the government (or possibly the SRO) enforces the contract $z(\cdot)$. Since the agent's payment, report, and the contract are all verifiable, commitment to a policy of conducting a (costly) investigation and taking all of the agent's wealth if the payment is less than $z(r)$ enforces compliance at no cost in equilibrium.

$$v(w,r|z,p,x) \equiv p(r)\, u(\max[w - z(r) - x(w,r),\, 0]) + (1-p(r))\, u(w - z(r)).$$

This reflects the fact that the maximum penalty the agent can pay is the residual, $w - z(r)$.

Since the agent chooses a report to maximize this expected utility, in equilibrium the reporting strategy must satisfy the incentive constraint

(AIC) $v(w,r(w)|z,p,x) \geq v(w,s|z,p,x)$ for all w and s such that $z(s) \leq w$.

2.2 The Customer's Problem

Having characterized the agent's problem, now consider the problem faced by the customer when choosing a contract z to offer the agent. Since the agent's reservation payoff is 0, and since any contract offers a non-negative payoff to the agent, the agent will accept any offer. Thus, the customer's problem is to choose the contract z that maximizes the customer's expected payoff taking into account the agent's optimal reporting strategy. Taking the enforcement policy (p,x) as given, the customer's problem can be written as the following mechanism design problem:

CP(p,x): $\max_{z,r}$ $E[\, z(r(W))\,]$

subject to (AF) $z(r(w)) \leq w$ for all w,

(AIC) $v(w,r(w)|z,p,x) \geq v(w,s|z,p,x)$ for all w,s such that $z(s) \leq w$.

Any equilibrium must satisfy the customer's incentive constraint corresponding to this subgame:

(CIC) (z,r) solves CP(p,x).

In addition to collecting the payments $z(r(W))$, the customer must also pay the transaction fee t to the SRO. Since the customer has an outside opportunity of α, the customer chooses to contract with the agent only if the customer's individual rationality constraint is satisfied:

(CIR) $E[\, z(r(W))\,] - t \geq \alpha$.

2.3 The SRO's Problem

The SRO chooses (p,x,t) to maximize the agent's expected utility. The transaction fee t must cover the SRO's expected investigation cost net of penalties. This yields the SRO's budget constraint:

(RB) $t \geq E[\, p(r(W))\, (c - \min[x(W,\, r(W)),\, W - z(r(W))])\,]$,

where $\min[x(W, r(W)), W - z(r(W))]$ is the actual penalty recovered given the agent's resource constraint.[8]

Given this constraint together with the characterization of the customer's and agent's optimal strategies, the SRO's problem can be written as the following mechanism design problem:

$$\text{SRP:} \qquad \max_{z,r,p,x,t} \quad E[\, v(W,r(W)|z,p,x) \,]$$

$$\text{subject to} \quad \text{(CIC)} \quad (z,r) \text{ solves } \text{CP}(p,x),$$

$$\text{(CIR)} \quad E[\, z(r(W)) \,] - t \geq \alpha,$$

$$\text{(RB)} \quad t \geq E[\, p(r(W)) \, (c - \min[x(W,r(W)), W - z(r(W))]) \,].$$

We have assumed that the SRO's objective is to maximize the agent's expected utility. Note, though, that the general approach would be the same if the enforcement policy and transaction fee were chosen by a regulator (say the government) whose objective put weight on both the agent's and customer's expected utility. In this case, the regulator chooses some point on the Pareto frontier. Hence, to characterize the solutions for this more general case, it suffices to solve SRP for alternative values of α.

Before analyzing self-regulation, it is useful to put the problem in perspective by comparing our approach to models of contracting in the existing literature. Previous analyses assume that one party, either the customer or the agent, chooses both the contract and the enforcement policy. With perfect competition among agents, and the customer choosing both the contract and the enforcement policy, the best the customer can do is solve the following problem:

$$\max_{z,r,p,x,t} E[z(r(W)) \,] - t \qquad (1)$$

subject to (AF), (AIC), and (RB).

If the solution to this problem yields the customer a payoff lower than α, then the customer will not participate. Alternatively, with a monopolistic agent choosing everything the problem can be written as:

$$\max_{z,r,p,x,t} E[\, v(W,r(W)|z,p,x) \,] \qquad (2)$$

subject to (AF), (AIC), (CIR) and (RB).

[8] The regulator balances the expected budget. This corresponds to the case in which the regulator oversees many such transactions and can rely on the Law of Large Numbers.

In a sense, SRP is intermediate to these two, with the agent choosing (p,x,t) and the customer choosing (z,r). Hence we might expect the agent's and customer's expected utility to be intermediate to that in (1) and (2). In fact, we will show that the customer may fare no better with self-regulation than with monopoly.

3. Optimal Contracts and Enforcement

We now characterize the solution to SRP. We show that the SRO's ability to choose the enforcement policy conveys substantial market power to the otherwise competitive agents. Even though the customer makes a take-it-or-leave-it contract offer to the agent, if the customer's reservation utility is not too low (specifically, $\alpha \geq \underline{w}$), the customer receives his reservation utility and the agent receives all of the rents.

To provide intuition for the solution we begin by considering the simpler problem in which the agent is a monopolist, as in (2). This is a standard mechanism design problem. Thus, the first step is to invoke the revelation principle to restrict attention to direct mechanisms in which the agent reports truthfully, $r(W) = W$. With truth-telling, the penalty $x(W,r)$ for $r \neq W$ matters for the agent's incentive constraint but is not incurred in equilibrium. So without loss of generality we can impose the maximum penalty and leave the agent with zero consumption if he is caught lying. What about the penalty $x(W,W)$ imposed when the agent tells the truth? If this penalty is positive, then in equilibrium the agent faces the risk of an investigation and penalty. Since the agent is risk averse, the agent is better off if we instead reduce his compensation (i.e., raise z) by the expected penalty, and impose no penalty when he tells the truth. As a final simplification, note that the transaction tax can be set as low as possible (to relax (CIR)) until the budget constraint (RB) binds; since no penalties are collected in equilibrium, t equals the expected investigation costs. The same results apply when the agents are perfectly competitive, as in (1). Summarizing, we have

PROPOSITION 1. The solution to the monopolist agent problem, (2), is characterized by

$$\text{MA:} \qquad \max_{z,p} \qquad E[\, u(W - z(W)) \,]$$

$$\text{subject to} \quad (\text{AF}^*) \quad z(w) \leq w,$$

(AIC*) $u(w-z(w)) \geq (1-p(w'))\, u(w - z(w'))$ for all w' such that $z(w') \leq w$,

(CIR*) $E[\, z(W) - p(W)\, c\,] \geq \alpha$,

with $r(w) = w$, $t = E[\, p(W)\, c\,]$, $x(w,w) = 0$ and $x(w,w') = w - z(w')$ for $w' \neq w$. The competitive solution, (1), is identical, with the objective replaced by $E[\, z(W) - p(W)\, c\,]$.

The same approach cannot be taken to solve SRP. The SRO's problem is not a standard mechanism design problem. Among its constraints is another mechanism design problem, the one involving the customer's contract choice. This nesting of the contract design implies that the revelation principle cannot be directly invoked in this case.

Instead, consider first the following simple implication of the customer's problem: the customer will always demand to be paid at least $\underline{w}$. This is the amount the customer knows the agent has for sure, independent of any reporting. Thus, at the solution to the customer's problem, $z(r) \geq \underline{w}$, and so SRP must satisfy this as well (see Lemma 1 in the Appendix). This observation implies that the SRO can do no better than the following problem,

SRP*: $\max_{z,p}$ $E[\, u(W - z(W))\,]$

subject to (AF*), (AIC*), (CIR*), and

(ZW*) $z(w) \geq \underline{w}$.

SRP* is the monopoly problem, MA, with the added constraint (ZW*) that the customer receive at least $\underline{w}$.

A second key observation is the following. In solving the monopoly problem MA, since the penalties are not paid in equilibrium, any penalties which deter the agent from misreporting are equivalent. Thus it is sufficient, but not necessary, to set maximal penalties when the agent misreports. However, in SRP, once the penalty is set the customer has the opportunity to propose the payment scheme z. The customer's ability to demand a high payment z will be restricted by the agent's incentive constraint – if z is too large, the agent will choose to misreport and suffer the consequences rather than pay z. Thus, in SRP the penalty schedule x plays two roles: to prevent the agent from misreporting, and to prevent the customer from asking for a higher payment z. This suggests that an optimal penalty schedule for the SRO

will do just enough to enforce the agent's reporting schedule, but no more. Any higher penalty will increase the customer's ability to raise z, to the agent's detriment.

To this end, given (z,p) we define the *weakest penalty* x^* as the smallest x such that the agent's incentive constraint (AIC*) exactly binds. That is, $x^*(w,w') = 0$ if $z(w') \geq z(w)$ and

$$u(w-z(w)) = (1-p(w'))u(w-z(w')) + p(w')u(w-z(w')-x^*(w,w')), \tag{3}$$

if $z(w') < z(w)$. Given x^*, if the customer proposes any increase in $z(w)$, the agent will lie. Thus, the penalty schedule x^* imposes the biggest limitation on the customer's choice of z while still supporting truth-telling.

Our main result is that the combination of the new constraint (ZW*) and the use of the weakest penalty schedule x^* characterizes the solution to SRP:

PROPOSITION 2. The solution to SRP is equivalent to the solution to SRP* with $r(w) = w$, $t = E[\, p(W)\, c\,]$, and $x = x^*$.

The result shows that the only equilibrium implication of the customer choosing the payment scheme is the constraint (ZW*). If the (ZW*) constraint does not bind for a monopolist, then the SRO's problem is equivalent to the agent having full monopoly power. The following proposition highlights this by examining the customer's equilibrium payoff.

PROPOSITION 3. If $\alpha \leq \underline{w}$, then the solution to SRP* is given by $z \equiv \underline{w}$ and $p \equiv 0$, and the customer's expected payoff is $\underline{w}$. In this case, monopoly is better than self-regulation for the agent. If $\alpha > \underline{w}$, then the customer's expected payoff is α and (CIR*) binds. If, in addition, the agent is risk neutral, self-regulation and monopoly coincide.

If $\alpha < \underline{w}$, the monopoly and self-regulator solutions differ since the optimal solution for a monopolistic agent is to offer the customer an expected payoff of α, while under self-regulation, the agent is forced to pay $\underline{w}$. If $\alpha \geq \underline{w}$, both a monopolist and a self-regulator would hold customers to their

reservation utility, α, but the solutions may differ; under the SRO the agent gets zero in the worst state, whereas a monopolist can choose $z(\underline{w}) < \underline{w}$ to provide consumption smoothing if the agent is risk averse.

If the agent is risk neutral, consumption smoothing is not an issue and the optimal contract will minimize enforcement costs. In particular, it reduces incentives to cheat by making z as "flat" as possible; i.e., by minimizing the difference between the highest and lowest payments. When $\alpha > \underline{w}$, this implies the monopolist will choose $z \geq \underline{w}$ (i.e., (ZW^*) does not bind) and the monopoly and SRO solutions coincide.

Next consider the optimal investigation policy.

PROPOSITION 4. At a solution (z,p) to SRP^*, the investigation probability p is weakly decreasing in z and is given by,

$$p^*(w) = \max_{z(w') \geq z(w)} 1 - [\, u(w' - z(w'))/u(w' - z(w)) \,].$$

If the agent is risk neutral, then at a solution to SRP the expected probability of an investigation, $E[\, p(W) \,]$, is weakly increasing in c and α.

The agent has the greatest incentive to cheat by reporting an outcome that requires a low payment z. Therefore these outcomes require the most vigorous enforcement and the investigation probability is decreasing in z. For the comparative statics, note that the investigation probability is increasing in the reservation utility of the customer, α, because the more the agent must pay the customer, the greater the agent's incentive to lie and report a low state. Surprisingly, the investigation probability also increases with the investigation cost, c. That is, the SRO's "demand function" for investigations is *upward* sloping. This follows since the agent's payment must cover the customer's reservation payoff and the expected enforcement cost. A higher investigation cost results in higher enforcement costs and therefore a higher expected payment from the agent. Thus, this has the same effect as an increase in α.

Finally, consider the relation between the optimal payment z and the outcome w. If it were feasible, it would be optimal to use a constant z since then the deadweight costs of enforcement could be avoided.

However, if $\alpha > \underline{w}$, such a scheme is infeasible since the agent will not be able to make the payment if the outcome is too low. This forces down the payments for low outcomes. Thus, it is natural to expect the optimal z to be increasing in w. Further, since the agent should be rewarded for reporting high outcomes, it is natural to expect the agent's payoff $w - z(w)$ to be increasing in w. In fact, under risk neutrality the agent receives an option-like contract.

PROPOSITION 5. If the agent is risk neutral, then the customer's payoff $z(w)$ is weakly increasing in w and the agent's payoff $w - z(w)$ is weakly increasing and convex in w.

3.1 Binary Outcomes

We illustrate self regulation for the binary case, $\Omega = \{w_1, w_2\}$, where $w_2 > \alpha > w_1$ (so $\underline{w} = w_1$). Let $z_i = z(w_i)$, $p_i = p(w_i)$ and $\pi_i = Pr(W = w_i)$ for $i = 1,2$. By (ZW*) and (AF*), we have $z_1 = w_1$. Using **PROPOSITION 4**, investigation probabilities are $p_1 = 1 - u(w_2 - z_2)/u(w_2 - w_1)$ and $p_2 = 0$. Using (3), the penalty for falsely reporting $W = w_1$ is $x(w_2,w_1) = w_2 - w_1$; in this case the weakest penalty x^* is the maximal penalty. There is no penalty for truthfully reporting $W = w_1$, $x(w_1,w_1) = 0$. Since $p_2 = 0$, other penalties are irrelevant. Since $\alpha > w_1$, the customer's individual rationality constraint, (CIR*), is binding, so $z_2 = [\alpha - \pi_1(w_1 - p_1 c)]/\pi_2$. The transaction fee equals the expected investigation cost, $t = \pi_1 p_1 c$.

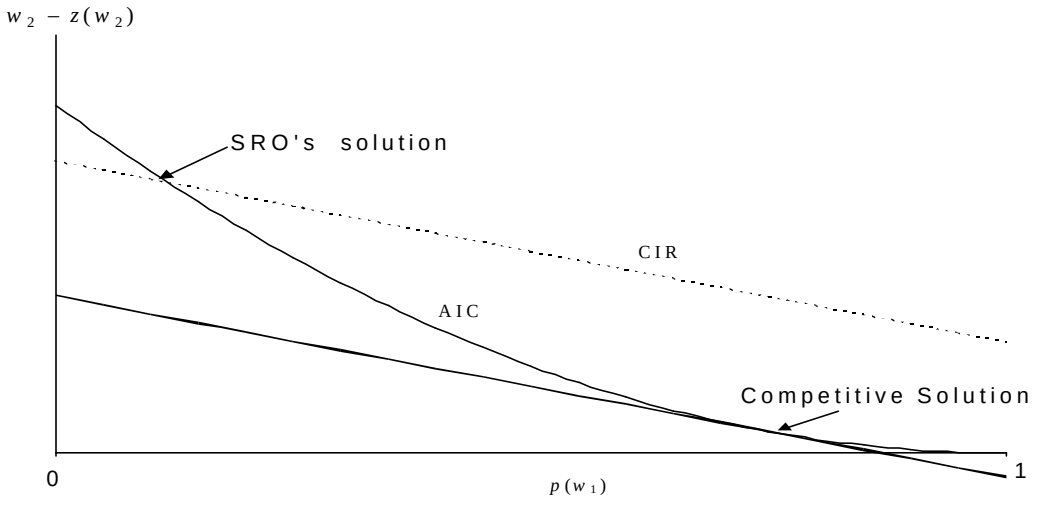

Figure 1. Competitive versus SRO contract.

Figure 1 shows the equilibrium choice of z_2 and p_1. Inducing truthful reporting of $W = w_2$ requires a high agent payoff $w_2 - z_2$ and/or a high probability p_1 of an investigation given a report of $W = w_1$. This is the region above (AIC*). The customer participates if his payoff z_2 is high enough and/or the investigation probability p_1 (and expected investigation cost) is low enough. This is the region below (CIR*).

Given the SRO's choice of p_1, the customer chooses the lowest $w_2 - z_2$ that induces the agent to report truthfully. The agent prefers to be induced to tell the truth via the "carrot" of a high payoff rather than via the "stick" of a high investigation probability. Hence the SRO prefers to set p_1 as low as possible and induce the customer to offer a high $w_2 - z_2$. If the required $w_2 - z_2$ is too high, however, the customer would not participate. The customer's individual-rationality constraint limits how low the SRO can set p_1. The intersection of (AIC*) and (CIR*) determines z_2 and p_1.

Figure 1 also shows the fully competitive outcome (the customer chooses everything) as in (1). The customer maximizes his expected payoff subject to the agent incentive-compatibility constraint, choosing the point of tangency between his indifference curve and (AIC*). Compared to the fully competitive outcome, self-regulation entails more lax enforcement — a lower investigation probability — and consequently a higher expected utility for the agent and a lower expected utility for the customer.

Next compare self-regulation to the monopolistic agent as in (2). If, as with self-regulation, $z_1 = w_1$ is chosen, then the optimal z_2 and p_1 are the same as in SRP. If, however, the monopolistic agent is risk averse and his marginal utility at zero consumption is sufficiently high, he will choose $z_1 < w_1$, ensuring positive consumption in both states. Figure 2 shows the effect of a monopolist choosing $z_1 < w_1$. The (AIC*) twists out; with z_1 lower and p_1 unchanged, reporting w_1 becomes more attractive so a higher agent payoff is required to induce truthful reporting of $W = w_2$. The (CIR*) shifts down; with z_1 lower and p_1 unchanged, a higher customer payoff when $W = w_2$ is required to induce customer participation. These changes in (AIC*) and (CIR*) imply that if the monopolistic agent chooses $z_1 < w_1$, he will also choose a higher investigation probability p_1 and a lower agent payoff $w_2 - z_2$ than will an SRO.

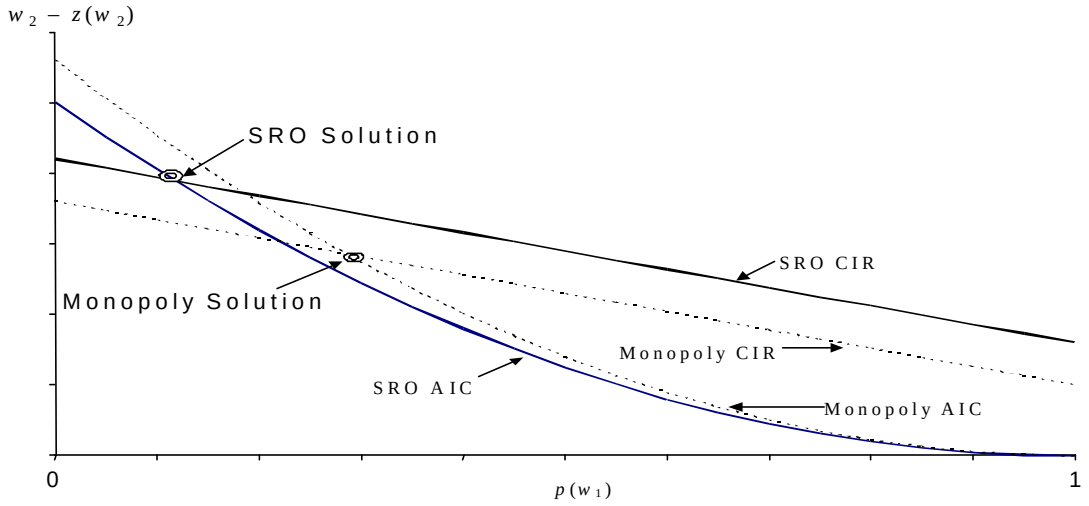

Figure 2. Monopoly versus SRO solution.

Finally, consider comparative statics. In the fully competitive case, an increase in the investigation cost c leads to a lower investigation probability p_1 and higher agent payoff $w_2 - z_2$. With self-regulation or monopoly, an increase in c leads to a *higher* p_1 and lower $w_2 - z_2$. This can be seen in Figure 1 and Figure 2. An increase in c or π_1 leads to steeper customer indifference curves (CIR*) (they pivot about the vertical axis intercept); an increase in p_1 is more costly so indifference requires a larger decrease in $w_2 - z_2$. With self-regulation or monopoly, this shift leads to a higher p_1 and a lower $w_2 - z_2$. By contrast, in the fully competitive case, this shift leads to a lower p_1 and a higher $w_2 - z_2$. Lastly, an increase in the customer's reservation utility, α, causes a downward parallel shift to (CIR*). This leads an SRO or monopolist to choose a higher p_1 and a lower $w_2 - z_2$.

3.2 A Numerical Example: Non-Maximal Penalties

With a binary outcome, the SRO's optimal enforcement policy entails the maximum penalty for fraud. As illustrated in the following example, this is not true in general. Suppose W is drawn uniformly from the set {100, 150, ..., 500}. Let $c = 100$, $\alpha = 200$ and let the agent be risk averse with utility $u(y) = y^{1/2}$. Then the optimal payment schedule $z(W)$ is illustrated in Figure 3, together with the investigation policy $p(W)$ (represented in % on the same scale).

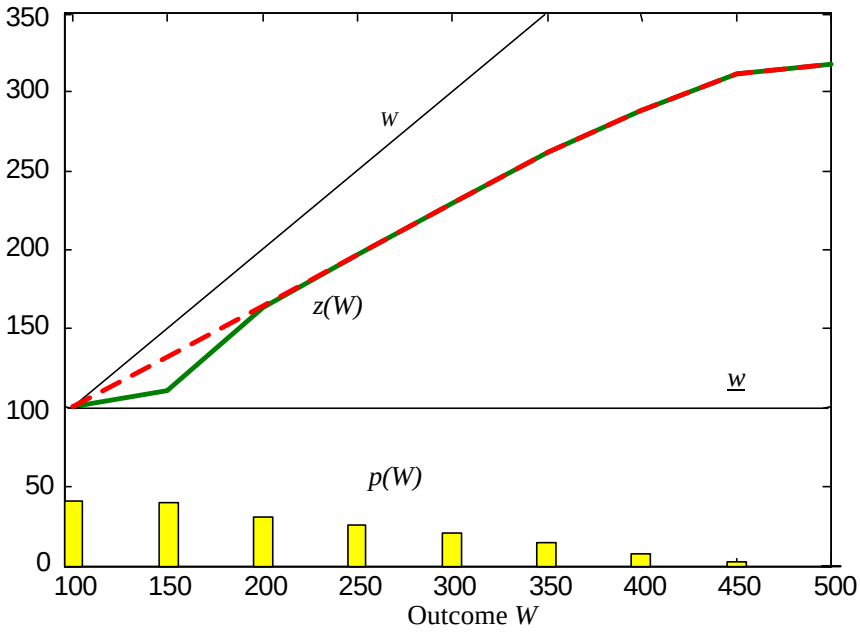

Figure 3. Multi-State Example with Non-Maximal Penalties

The dashed line represents the solution to the customer's problem *if the maximal penalty were imposed*. Where z is below the dashed line, the SRO's penalties x^* are below the maximum. In this example, if the agent is caught reporting 100 when the outcome is 150, the SRO imposes a fine of 21.1 rather than the maximal fine of 50. This prevents the customer from demanding a higher z when the outcome is 150. Intuitively, the SRO uses weak penalties when possible to shift rents to the agent in those states for which the agent's marginal utility is high.[9]

3.3 Heterogeneous Customers

Thus far, we have assumed that all customers have the same opportunity cost. In this section we allow this opportunity cost to be heterogeneous. Let α_i denote customer i's opportunity cost and let $F(\alpha)$ represent the fraction of the customer population with opportunity cost below α. In this case, the enforcement policy chosen by the SRO will determine the fraction of the population that is willing to

[9] By contrast, in some settings the role of non-maximal penalties is to induce malfeasants to commit less serious rather than more serious crimes. This motivation is not present here, since all fraud is deterred.

transact with an agent. The SRO thus faces a tradeoff between the expected rents of an agent per transaction and the volume of trade.

Once the SRO sets an enforcement policy (p,x,t), competition between agents implies that each customer i will negotiate the (z,r) that solves CP(p,x) (this solution is independent of α_i). This yields an expected payoff of $E[z(r(W)) - t]$. Thus, only customers with $\alpha_i \leq \alpha \equiv E[z(r(W)) - t]$ will transact with the agent, resulting in volume of $F(\alpha)$. Therefore, the SRO's problem can be decomposed as follows. First, the SRO chooses an α determining the fraction of the customer population that will be served. Then the enforcement policy is chosen as the solution to SRP given α.

The highest feasible α that the SRO can choose is the solution to the fully competitive case, defined by (1); denote this by α^C. The minimum α that the SRO can choose is $\underline{w}$. This follows from **PROPOSITION 3**. So if we denote by $V(\alpha)$ the agent's expected utility at the solution to SRP given α, then the SRO chooses α to solve

$$\max_{\alpha \in [\underline{w}, \alpha^C]} F(\alpha)V(\alpha). \qquad (4)$$

Denote the solution to (4) as α^{SRO}. Clearly this solution yields a lower expected customer payoff than the fully competitive case. Interestingly, α^{SRO} may be higher or lower than the solution to the pure monopolist case, given by $\max_\alpha F(\alpha)\, V^M(\alpha)$, where $V^M(\alpha)$ is the agent's expected utility at the solution to (2) given α. So customers could be better off dealing with a pure monopolist as compared to an SRO.[10]

4. Government Oversight of the SRO

We now consider government oversight of SRO enforcement. Oversight is an additional tier on the contracting problem where the government can also investigate the agent and impose penalties. We show that the threat of government enforcement is sufficient to induce greater enforcement by the SRO. In

[10] We established this result with a numerical example in which the agent becomes very risk averse at low levels of consumption. In this case, the pure monopolist can attain much higher utility than the SRO since the pure monopolist can ignore the (ZW*) constraint – thus giving the agent a positive payoff in all states. This induces the pure monopolist to offer higher customer utility in order to transact with a larger fraction of the customer population.

equilibrium, the SRO makes its enforcement just aggressive enough to keep the government from doing any enforcement of its own.

The essence of oversight is that the government can conduct an investigation if the SRO does not. Presumably an SRO has greater expertise and better information than government regulators and so can more easily determine whether a member has cheated. So we assume that the government's cost of investigating the agent, denoted c_g, is higher than the SRO's cost, $c_g \geq c$.[11] Besides conducting investigations, the government can impose its own monetary penalties beyond any imposed by the SRO, or can reduce a penalty imposed by the SRO.[12] We assume that the government's objective is to maximize customer expected utility.

After a report r by the agent, let $p_s(r)$ and $p_g(r)$ denote the probability of an investigation by the SRO and the government, respectively. We assume that the SRO decides first whether to investigate. Since investigations are costly and observable, the government will never conduct a redundant investigation. Thus, the probability that the agent is investigated given that he reports r is given by $p(r) \equiv p_s(r)+p_g(r)$.[13]

Since the government can raise or lower the penalties imposed by the SRO, given penalties x_s set by the SRO and x_g set by the government, the effective penalty of the agent is simply x_g. Of this amount, if the SRO investigates, $\min[x_s, x_g]$ is collected by the SRO and $\max[x_g - x_s, 0]$ is collected by the government. If the government investigates, then x_g is collected by the government.

[11] This assumption justifies the existence of the SRO. If the government's investigation cost were lower than the SRO's, it would be Pareto optimal to eliminate the SRO and have all investigations performed by the government.

[12] For example, the SEC, state securities regulators, a state attorney general, or the Justice Department can initiate its own enforcement action and impose its own penalty. With regard to a reduction in SRO penalties, an SRO sanction is subject to review first by the SEC and then by the federal court of appeals. The SEC or court can reduce or cancel a sanction. See Frankhauser, et al. (1997) and Phillips (1997). We continue to limit the analysis to monetary penalties. A role of government that we do not consider is the imposition of non-monetary penalties, like jail.

[13] This is equivalent to the following. The SRO investigates with probability $p_s(r)$. If the SRO does not investigate, the government investigates with probability $p' = p_g(r)/(1-p_s(r))$. The total investigation probability is $p(r) = p_s(r) + (1-p_s(r))p' = p_s(r) + p_g(r)$ and the unconditional probability of a government investigation is $(1-p_s(r))p' = p_g(r)$.

The transaction fees used to fund the SRO's and government's enforcement activities are denoted t_s and t_g, respectively. The customer thus pays $t \equiv t_s+t_g$ to participate in this market. Finally, we incorporate heterogeneous customers as in Section 3.3.

We specify the timing of our model with government oversight to reflect the idea that the government can respond to the SRO's enforcement policy. Specifically, the timing is as follows:

1. The SRO chooses an enforcement policy (p_s, x_s) and a transaction fee t_s.
2. Taking the SRO's enforcement policy and transaction fee, (p_s,x_s,t_s), as given, the government chooses its enforcement policy (p_g, x_g) and its transaction fee t_g.
3. Taking the overall enforcement policy, (p_s+p_g,x_g,t_s+t_g), as given, customers choose whether to participate. Each customer offers a contract z to the agent, which the agent can accept or reject.
4. If the agent rejects the contract, the agent receives 0 and the customer receives his reservation utility. If the agent accepts the contract, the customer pays t_s+t_g and the problem continues.
5. A cash flow w is realized. Only the agent observes this realization.
6. The agent chooses a cash flow, r, to report and gives the customer the corresponding payoff, $z(r)$.
7. Given the reported cash flow r, the agent is investigated with probability $p(r)$. If the SRO investigates the agent, the SRO pays c. If the government investigates the agent, the government pays c_g. In either case, the agent pays the penalty $x_g(w,r)$.

The customer's problem and the agent's problem are the same as before. Taking (p,x_g) as given, the customer's problem is given by CP(p,x_g).

4.1 Analysis of Government Oversight

The government takes the SRO's enforcement policy as given and chooses its own enforcement policy to maximize the customer's expected utility. A full statement of the government's mechanism design problem appears in the Appendix. Here, we simplify the analysis by focusing on the binary case,

$\Omega = \{w_1, w_2\}$. In this case, investigations are only useful if the agent reports a low outcome (see section 3.1). Thus we assume that $p_s(w) = 0$ for $w \neq w_1$, and consider the SRO's choice of $p_s(w_1)$.[14]

Since the government shares the customer's objective, this is much like a standard mechanism design problem in which the government chooses the entire contract. This leads to the following, simplified representation of the government's problem:

PROPOSITION 6. The government's problem is equivalent to

$$\begin{aligned} \text{GP}^*(p_s): \quad & \max_{z,p,t_g} \quad E[\, z(W)\,] - t_g \\ \text{subject to} \quad & (\text{AF}^*) \quad z(w) \le w \text{ for all } w, \\ & (\text{AIC}^*) \quad u(w_2 - z(w_2)) \ge (1 - p(w_1))\, u(w_2 - z(w_1)), \\ & (\text{GB}^*) \quad t_g \ge E[\, (p(W) - p_s(W))^+ \,]\, c_g, \end{aligned}$$

Notice that p is used to represent the aggregate investigation probability, (AIC*) is written for the binary case, and maximal penalties are used if the agent falsely reports w_1 and is investigated.

Anticipating government oversight, the SRO chooses an enforcement policy to maximize the agent's expected utility. In stating the SRO's problem, note that the penalties assigned by the SRO play no role in equilibrium (they are overridden by the government). Second, the tax t_s is set to cover the expected investigation costs of the SRO. This leads to the following SRO problem with oversight:

$$\begin{aligned} \text{SRPO:} \quad & \max_{p_s,z,p,\alpha,t_s,t_g} \quad F(\alpha) E[\, u(W - z(W))\,] \\ \text{subject to} \quad & (\text{CIR}') \quad E[\, z(W)\,] - t_s - t_g \ge \alpha, \\ & (\text{RB}') \quad t_s \ge E[\, p_s(W)\,]\, c, \\ & (\text{GIC}) \quad (z,p,t_g) \text{ solves } \text{GP}^*(p_s). \end{aligned}$$

[14] This restriction is without loss of generality in the binary case. In a more general case, we would need to explicitly incorporate the reporting strategy into the government's problem. The revelation principle would not apply since the government could use r to exploit the SRO's investigation policy.

The SRO maximizes the agent's expected utility per customer multiplied by the fraction of the customer population that chooses to participate. Besides incorporating heterogeneous customers, the key difference between SRP and SRPO is that the latter has the constraint of government oversight, (GIC).

Since the SRO prefers less enforcement than the government, one might expect the SRO to do no investigations. However, as the following result shows, at the solution to SRPO the SRO does all investigations – the government conducts none.

PROPOSITION 7. Let p^{cg} be the investigation probability at the solution with perfect competition, (1), given investigation cost c_g. Let p^{srp} be the investigation probability at a solution to SRP with heterogeneous agents and investigation cost c, as in (4). Then at the solution to SRPO, $p_g = 0$ and $p_s \geq \max[\, p^{srp}, p^{cg}\,]$, with equality if $F(\alpha)$ is log-concave or $p^{srp} \geq p^{cg}$. [15]

Thus, with government oversight, no actual enforcement by the government is observed. But if $p^{srp} < p^{cg}$, government oversight is effective and increases the customer's expected payoff. If $p^{srp} \geq p^{cg}$, oversight has no effect (this is the case if government investigation is very costly). To see the intuition, note that for any investigation probability chosen by the SRO, the government will augment it with its own investigation probability as long as it is efficient to do so; that is up to p^{cg}. Given this, the SRO will pre-empt the government since it's investigation cost is lower than the government's, and this lower cost leads to higher customer participation.

In equilibrium, oversight entails no government enforcement. But oversight does require a credible threat of enforcement and this is likely to entail a fixed cost. For instance, the SEC will incur costs to maintain the staff needed for oversight. Consequently employing government oversight is optimal if this fixed cost is low, or if the marginal enforcement cost, c_g, is low (since this cost determines the gain from oversight even though it is not incurred in equilibrium).

[15] Note that since z follows from p from (AIC*), it immediately follows that $z \geq \max[\, z^{srp}, z^{cg}\,]$ with equality if $F(\alpha)$ is log-concave or $p^{srp} \geq p^{cg}$, where z^{srp} and z^{cg} have the corresponding definitions.

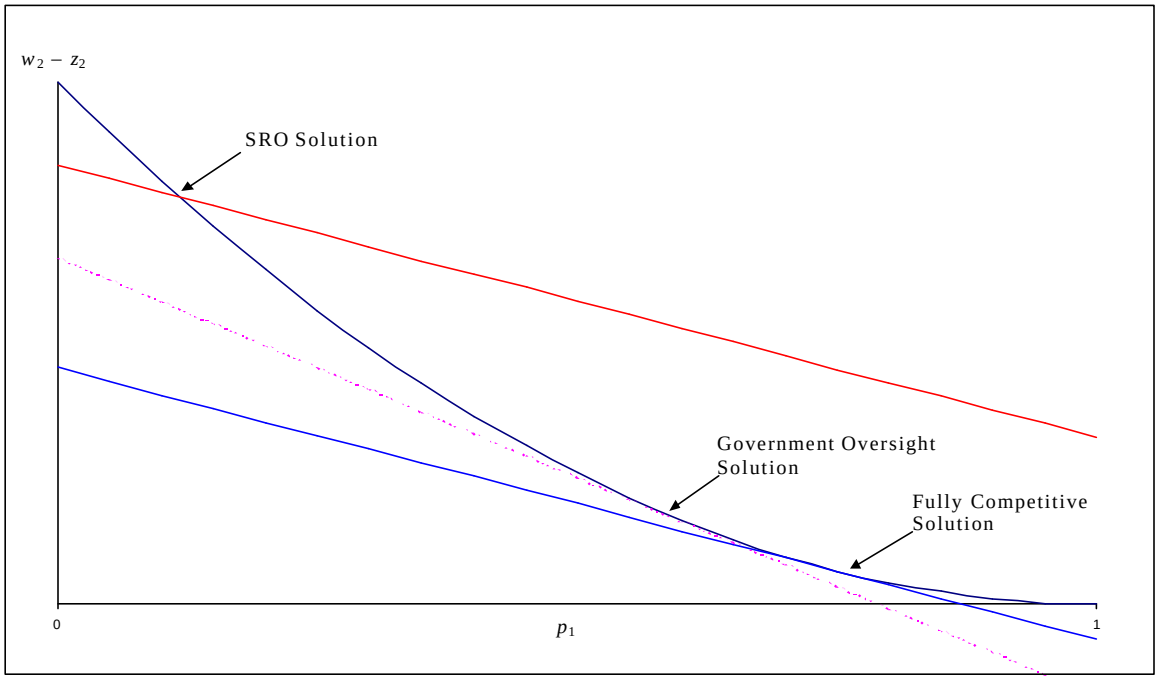

Figure 4. Government Oversight

Figure 4 shows the solution with government oversight. It is the tangency between the customer's indifference curve given an investigation cost c_g (the government's) and the agent's incentive compatibility constraint. If the SRO were to choose a lower investigation probability, the government can raise the customer's expected payoff by doing its own investigations, pushing the total investigation probability up to the point of this tangency. With $c_g > c$, the resulting solution, though preferred to no oversight, still entails less enforcement than the fully competitive case. The fully competitive solution is at the tangency between the customer's indifference curve given an investigation cost c (the SRO's) and the agent's incentive compatibility constraint.

4.2 Further Remarks on Government Oversight

Earlier we showed that contrary to customers' interests, the SRO may impose a less than maximal penalty for fraud. This is why the government's ability to impose a penalty may benefit customers. With oversight, the government's ability to reduce an SRO penalty is helpful too. Without it, if the agent is risk averse, the SRO may impose penalties even when an investigation reveals that there is no fraud. By doing so, the SRO makes the agent's payoff random, raising his marginal utility of income and making it more

costly for the government to raise z via additional enforcement.[16] Consequently the government's ability to reduce an SRO penalty may benefit customers as well.

We assumed that the government observes which agents the SRO investigates. Then the government focuses its own investigations, if any, on agents not already investigated. Thus it costs the government $E[\ p_g(W)\ ]\ c_g$ to raise the investigation likelihood by p_g. Suppose, though, that the government cannot observe whom the SRO investigated. While the government knows the SRO did investigate the fraction $p_s(w_1)$ of agents, it does not know which ones. [17] If so, the fraction $p_s(w_1)$ of the government's investigations will be redundant. In this case, raising the investigation likelihood by p_g costs the government $E[\ p_g(W)/(1 - p_s(r(W)))\ ]\ c_g$. This marginal cost is increasing in the frequency with which the SRO investigates. In effect, SRO investigations would pre-empt government investigations by raising the cost. In the two-state case, the SRO will choose $p_s(w_1)$ at least up to the competitive solution with investigation cost of $c_g / (1 - p_s(w_1))$. Indeed, pre-emption by the SRO could occur even if $c_g < c$.

5. Reputation in a Dynamic Setting

Now we extend our results to a dynamic setting. Here the SRO can also penalize the agent by restricting his ability to transact with customers in the future. In addition, customers can penalize agents by choosing to not transact with them in the future, even if there has been no fraud. Since agents earn positive expected rents per transaction, total penalties can now be increased to include the discounted value of the anticipated future rents.

---

[16] To see this in the binary case, suppose an agent who reports $W = w_2$ (clearly no fraud) may be investigated and penalized. Letting $z_2 = z(w_2)$, $x_2 = x(w_2, w_2)$, and $p_i = p(w_i)$, the agent incentive compatibility constraint is $p_2 u(w_2 - z_2 - x_2) + (1 - p_2)u(w_2 - z_2) = (1 - p_1)u(w_2 - w_1)$. A higher customer payoff z_2 requires a higher investigation probability p_1. The marginal benefit of increasing p_1 is $\partial z_2 / \partial p_1 = u(w_2 - w_1)/[p_2 u'(w_2 - z_2 - x_2) + (1 - p_2)u'(w_2 - z_2)]$. With a risk-averse agent, this marginal benefit is lower if $x_2 > 0$ (compared to $x_2 = 0$). The penalty makes it more costly to satisfy the incentive compatibility constraint and thus more costly to induce a higher payment z_2.

[17] This is especially natural under the interpretation given in footnote 5. There, all agents reporting low outcomes appear to have been investigated.

We show that when investigation costs are sufficiently high, it is optimal for the SRO to do no enforcement. Instead, customers monitor agents' reputations and no longer transact with an agent who has reported a low outcome. This threat induces truth-telling by the agent. If investigation costs are low, the SRO investigates with positive probability and an agent who is caught cheating is penalized and barred from transacting in the future. In this case, the SRO chooses an enforcement policy that eliminates customers' incentive to monitor agents' reputations. Interestingly, the agent is motivated to tell the truth through either the threat of an investigation or the threat that his reputation will suffer, but not both.

We analyze an infinitely repeated version of our basic one-period SRO model (for simplicity we do not include heterogeneous customers or government oversight here). Each period a new customer arrives to transact. The customer chooses from among an infinite number of agents; in particular, the customer can choose the agent who handled last period's transaction or the customer can choose a new agent. Then the problem proceeds as before.[18]

Agents discount future utility according to the discount factor $\delta < 1$. Let $V^{\infty}(\alpha)$ denote the highest possible discounted expected utility of the agent given an expected customer payoff of at least α each period. Given our normalization $u(0) = 0$, the lowest possible utility for an agent is 0, which corresponds to a permanent suspension from transacting. Thus, after any history, the agent must receive a continuation utility between 0 and $V^{\infty}(\alpha)$, which can always be implemented by randomizing between permanent suspension and no penalty. Hence, rather than consider all possible intermediate punishments (such as temporary suspensions, fines to be repaid from future earnings, etc.), we can without loss of generality suppose that the only punishment is permanent suspension, applied stochastically.

To see how this alters the general formulation, let $\phi(w,r)$ be the probability of suspension if the agent is investigated and reported r when the true outcome is w. Let $\theta(r)$ be the probability of suspension if the agent reports r and is not investigated. The agent's expected utility given outcome w and report r is

[18] Our model features a sequence of one-period contracts. DeMarzo and Fishman (2000) characterize optimal long-term contracts when the agent privately observes the cash flows but there is no investigation technology.

$$v(w,r|z,p,x,\phi,\theta) = p(r)[u(\max[w-z(r)-x(w,r),0]) + \delta\ (1-\phi(w,r))\ V^{\infty}(\alpha)]$$

$$+ (1-p(r))[u(w-z(r)) + \delta\ (1-\theta(r))\ V^{\infty}(\alpha)].$$

Given this definition of v, we can define the SRO's problem as in Section 2.3. For now, consider the case in which only the SRO chooses $(p,\ x,\ \phi,\ \theta)$. In Section 5.1, we consider the consequences of customers' ability to increase θ by monitoring an agent's reputation.

Solving the SRO's problem implies a discounted utility for the agent, which in equilibrium equals $V^{\infty}(\alpha)$. This allows us to solve for $V^{\infty}(\alpha)$ using a dynamic programming approach.

We simplify the model to two states, $W \in \{w_1,\ w_2\}$ with $w_1 < \alpha < w_2$, with risk-neutral agents. In this case, as in Section 3.1, the optimal policy involves truth-telling. If there is an investigation, it is optimal to impose maximal penalties if the agent is caught cheating and no penalty otherwise; that is, $x(w_2,\ w_1) = w_2$, $\phi(w_2,\ w_1) = 1$, and otherwise x and ϕ equal 0. If there is no investigation, the agent is suspended according to θ. Thus, the overall probability of suspension given a truthful report is $(1 - p(w))\ \theta(w)$. This leads to the following specification for SRP$^{\infty}$:

SRP$^{\infty}$: $\quad V^{\infty}(\alpha) = \max_{z,p,\theta}\ E[\ W - \ z(W) + \delta\ (1 - (1 - p(W))\ \theta(W))\ V^{\infty}(\alpha)\ ]$

subject to

(AF*) $\quad z(w) \le w$

(AIC$^{\infty}$) $\quad w - z(w) + \delta\ (1 - (1 - p(W))\ \theta(W))\ V^{\infty}(\alpha)$

$$\ge (1 - p(w'))\ (w - z(w') + \delta\ (1 - \theta(w'))\ V^{\infty}(\alpha))$$

for all w' such that $z(w') \le w$

(ZW*) $\quad z(w) \ge \underline{w}$

(CIR*) $\quad E[\ z(W) - p(W)\ c\ ] \ge \alpha$

The following result is immediate – if the agent reports the high payoff, no investigation or suspension is needed. In the two-state case, the only relevant (and binding) agent incentive constraint prevents an agent with a high payoff from reporting a low one. Letting $z_j = z(w_j)$, $p_j = p(w_j)$, $\theta_j = \theta(w_j)$, and $\pi_j = Pr(W=w_j)$ we can write α and V^{∞} as functions of p_1 and θ_1 and we have

PROPOSITION 8. At a solution to SRP$^{\infty}$, $z_2 \geq z_1 = w_1$ and $p_2 = \theta_2 = 0$. Also, both (AIC$^{\infty}$) and (CIR*) bind, and

$$V^{\infty} = \frac{\pi_2(1-p_1)(w_2-w_1)}{1-\delta\left[(1-p_1)(1-\theta_1)+\pi_1 p_1\right]} \tag{5}$$

$$\alpha = \pi_2[\, E[W] - (1-\delta)\, V^{\infty}\,] + \pi_1\, w_1 + \pi_2\, \pi_1\, p_1\, [\, w_2 - w_1 + \delta\, V^{\infty}\,] - \pi_1\, p_1\, c. \tag{6}$$

This result makes clear that the SRO has two instruments, p_1 and θ_1, with which to provide incentives. Raising p_1 reduces the agent's payoff from cheating, but entails an investigation cost. Raising θ_1 also reduces the payoff from cheating, and has no direct cost. However, $\theta_1 > 0$ implies that the agent may be suspended for honestly reporting a low outcome. This reduces the agent's future rents. In essence, p_1 is precise but directly costly, whereas θ_1 is imprecise and thus indirectly costly. Of interest is how an SRO optimally trades off these two instruments.

Our main result is that optimal enforcement in a dynamic setting is characterized by distinct regimes in which one of the enforcement mechanisms is at a boundary. In an "investigation" regime, $\theta_1 = 0$ and an agent is only suspended if caught cheating. In a "performance-based" regime, $p_1 = 0$ and an agent may be suspended simply for reporting a low outcome. In a "zero-tolerance" regime, $\theta_1 = 1$ and an agent who reports a low outcome is suspended unless he is investigated and cleared of fraud. Notably, it is never optimal to mix both enforcement mechanisms, choosing both θ_1 and p_1 interior. Specifically, we have

PROPOSITION 9. At a solution to SRP$^{\infty}$, there are α^0, α^1 with $\alpha^0 < \alpha^1$ such that (i) for $\alpha \in (\underline{w}, \alpha^0]$, agents are only suspended after an investigation, $p_1 > 0$ and $\theta_1 = 0$; (ii) for $\alpha \in (\alpha^0, \alpha^1]$, agents who report low outcomes may be suspended and no investigations are conducted, $p_1 = 0$ and $\theta_1 > 0$; and (iii) for $\alpha > \alpha^1$, agents who report low outcomes are either investigated or suspended without investigation, $p_1 > 0$ and $\theta_1 = 1$. Moreover, α^0 is decreasing in c.

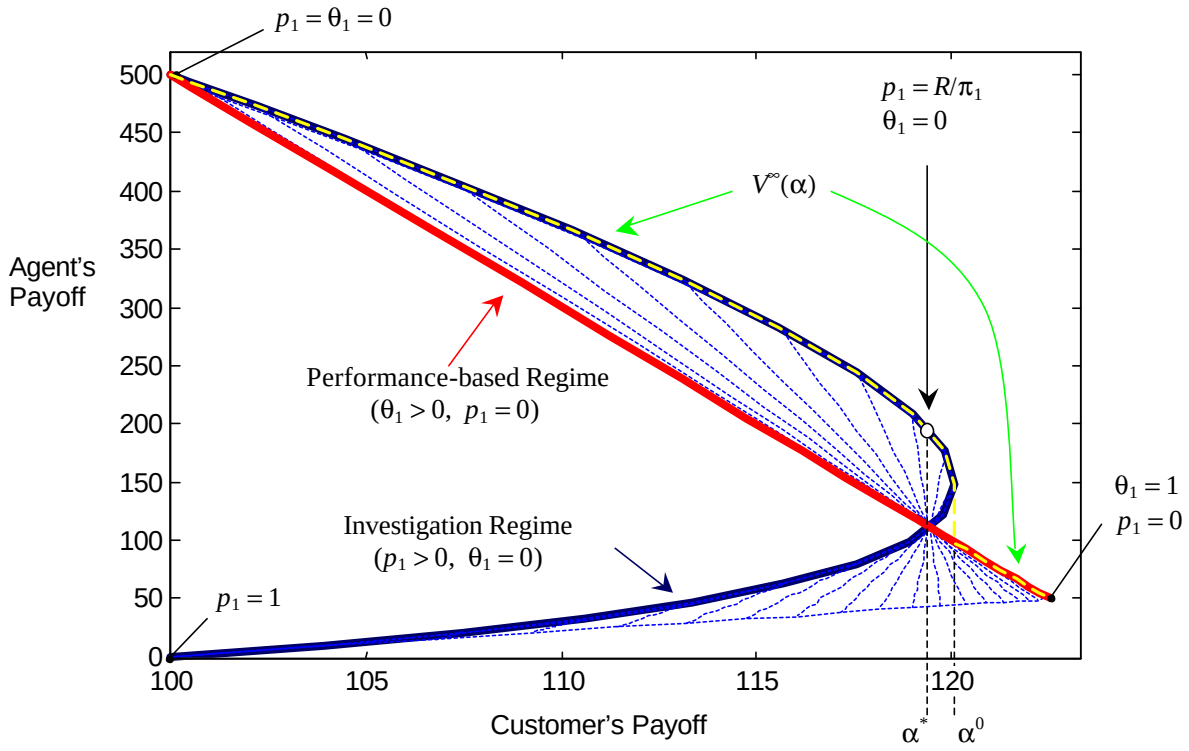

Figure 5. Payoff Possibilities for all (p_1, θ_1)

For example, suppose W is uniform on {100, 200}, the discount factor $\delta = .9$, and the investigation cost $c = 100$. Figure 5 plots the agent's and customer's utility for all combinations of p_1 and θ_1. The investigation regime corresponds to the parabola-shaped segment, with p_1 increasing clockwise along it. If this were a static example, with risk neutrality the customer's payoff would be monotone in p_1. This is not true in the dynamic case since the presence of future agent rents magnifies incentives. These rents are decreasing in p_1. So while initial increases in p_1 help the customer, further increases lower the rents and diminish incentives.

The performance-based regime corresponds to the linear segment, with θ_1 increasing to the right along it. The shaded region corresponds to interior (suboptimal) policies. In this example, the zero-tolerance regime is never optimal ($\alpha^1 = \infty$), which is true for all but the most extreme parameter values we considered.

The Pareto frontier V^{∞} is discontinuous, lying along the parabola for $\alpha \leq \alpha^0$ and jumping to the straight line for $\alpha > \alpha^0$. As the investigation cost c declines, the parabola and thus α^0 shifts out while the linear segment is unchanged. So a small improvement in investigation technology can lead to a switch from the performance-based regime to the investigation regime.

5.1 Customer "Oversight" and Agent Reputation

Thus far, we have considered an optimal policy under the assumption that both p_1 and θ_1 are chosen by the SRO. If prior outcomes are observable, however, customers can decide whether to use the current agent or switch to a new agent based on an agent's outcome history, or "reputation." Effectively, this allows customers to increase θ_1. Since customers cannot perform investigations, p_1 is still determined by the SRO.

We analyze the customers' optimal choice of θ_1 in this section. We maintain the previous results regarding x, ϕ, p_2 and θ_2; these choices are also optimal in this setting. Note that an individual customer's choice of whether to remain with the current agent or choose a new agent affects the incentives of the agent when dealing with the *previous* customer. That is, it does not benefit the current customer directly. The choice of θ_1 is therefore best interpreted as a "social norm" that specifies the extent to which customers use reputation when selecting an agent.[19]

[19] Clearly, there are multiple equilibria. In one, customers ignore past outcomes and choose the agent who has been in business longest. This leads to $\theta_1 = 0$. Moreover, in this equilibrium, if $p_1 > 0$ switching agents is suboptimal for a customer since the new agent would not expect to have future customers and so would face less of a threat if caught cheating. Another equilibrium is for customers to use the agent who has been in business longest without reporting a low outcome. This implements $\theta_1 = 1$. Again, following this rule is optimal for each customer since an agent who reported a low outcome in the past expects no future customers. Presumably, evolutionary forces will lead to the

In the example above, if $\alpha = 100$, $p_1 = \theta_1 = 0$ maximizes the agent's payoff. If, however, the SRO chose $p_1 = 0$, customers would have an incentive to not use any agent who has reported a low outcome in the past. This implies $\theta_1 = 1$, which leads to the highest payoff for the customer, and the lowest Pareto efficient payoff for the agent.

Given any policy (p_1, θ_1) of the SRO, it is natural to assume that customers choose a reputation-based strategy to increase θ_1 to maximize the customers' payoff. Starting from any p_1 in the investigation regime (the parabola) increasing θ_1 moves the payoffs along a line through the intersection of the investigation regime and the performance-based regime; see the dotted lines in Figure 5. From the figure, the customer's payoff is increasing in θ_1 if p_1 is low (the upper left region of the parabola), but is decreasing if p_1 is high. The "cutoff" level of p_1 is given below:

PROPOSITION 10. The agent's payoff is decreasing in θ_1. The customer's payoff is increasing in θ_1 if $p_1 < R / \pi_1$ and decreasing otherwise, where $R = (1-\delta)/\delta$ is the interest rate.

Thus, if customers can increase θ_1, they will do so unless $p_1 \geq R / \pi_1$. This limits the effective choice of the SRO to policies with $\theta_1 = 1$ or $p_1 \geq R / \pi_1$. In Figure 5, the SRO is effectively restricted to points beyond α^*, the customer's payoff at the intersection of the regimes. Thus, as with government oversight, if $\alpha < \alpha^*$ the customers implicit threat to use reputation increases monitoring by the SRO.

The intuition for this result is as follows. By increasing θ_1, customers increase the threat facing an agent who reports w_1, relaxing the (AIC$^\infty$) constraint and allowing them to demand higher z_2. However, increasing θ_1 also decreases V^∞, because it increases the likelihood that an honest agent will lose future business. This lowers the penalty when w_1 is reported, and so tightens the (AIC$^\infty$) constraint. This second effect dominates when R is low and the future rents are more important, when π_1 is high and honest

---

equilibrium that is best for customers. We do not model equilibrium selection here; we simply identify which equilibrium is best.

agents must report low outcomes more frequently, and when p_1 is high and the incentive to report truthfully is already strong.

An implication of this discussion is that in a dynamic setting, SRO enforcement may be more aggressive than is preferred by customers. In the example, the SRO would set $p_1 = R / \pi_1$. In response, customers would choose $\theta_1 = 0$ and resulting customer utility is α^*. Customers would do better with no enforcement, $p_1 = 0$, and strictly rely on reputation, $\theta_1 = 1$. So in this case, customers would be better off if they could choose to use agents who are not SRO members and who face no threat of investigation/penalty. In practice, securities industry customers do not have this choice; e.g., all brokers and dealers must be members of the NASD.

6. Concluding Remarks

A central message of this article is that through its control of the antifraud enforcement policy, an SRO can confer a great deal of market power to its otherwise competitive members. In general it does so by choosing a more lax enforcement policy than would be preferred by customers.

Prior to 1975, the NYSE fixed brokerage commissions by setting minimum allowable commissions. In effect, the exchange had full monopoly power, as in (2). Then the SEC (responding in part to pressure from the Department of Justice) eliminated minimum commissions. Since 1975 the exchange is no longer permitted to control commissions but still controls the enforcement policy, as in our model (see Jarrell (1984)). Our analysis suggests that even without the ability to fix brokerage commissions, the NYSE can still offer its members a degree of market power through its enforcement policy.

In the U.S., there are multiple SROs governing securities markets—in addition to the NASD and the NYSE, the regional stock and options exchanges are also SROs. In 1999, when drafting the Securities Markets Enhancement Act, the Senate Banking Committee proposed having the SEC designate a primary SRO that would examine broker-dealer firms and enforce the rules. One could build on our analysis to compare a single SRO with multiple competing SROs. Following the analysis in Section 3.3, suppose customers are heterogeneous and individual customers choose among multiple SRO-governed exchanges.

The SROs would compete through their enforcement policies, each SRO solving a problem like (4) with $F(\alpha)$ replaced by the fraction of customers served given the enforcement policies chosen by the various SROs. With a model like this, if exchanges are not perfect substitutes (in practice, exchanges trade different securities and have different trading protocols), one can show that enforcement and the resulting customer payoffs are increasing in the number of SROs (such a model appears in an earlier version of this article). This analysis suggests that even if brokers and dealers compete with one another, there is still a benefit to have multiple SROs that compete as well (due in part to opposition from state securities regulators, this item was not included in the draft legislation).

In our analysis, government oversight of self regulation can benefit customers by leading the SRO to engage in more aggressive enforcement. The SRO would choose an enforcement policy that is just aggressive enough to pre-empt the government from doing its own enforcement. This type of response is currently under way in accounting with regard to auditor independence. Due to concerns of conflicts of interest, the SEC has proposed rules to limit the consulting services that auditors can provide for clients and has proposed the establishment of a statutory regulator. In response, the AICPA established a new charter for its regulatory arm, the Public Oversight Board (POB). The new charter (dated 2/9/01, http://www.publicoversightboard.org/charter.htm) strengthens and broadens the POB's regulatory power, including raising its budget from $2 million to $5.2 million.

We also show that the possibility that customers can monitor agents' reputations over time can also lead the SRO to engage in more aggressive enforcement. As was the case with government oversight, the SRO may choose an enforcement policy that is just aggressive enough to pre-empt customers from relying on agents' reputations. The dynamic analysis also highlights that mandatory membership in an SRO can be beneficial for agents. In the one-period model, no agent would operate outside an SRO since no customer would do business with them without enforcement. In the dynamic model, however, there are circumstances under which customers would do business with non-SRO agents, relying solely on agent reputation to discipline agents. A government mandate requiring SRO membership would eliminate this

possibility and preserve agent rents. In the securities industry such a mandate is provided by the Securities and Exchange Act of 1934. Similar mandates exist in law, medicine and other professions.

Our analysis shows that lax enforcement is a means to enhance the rents of agents who are in a position to commit fraud. This implies that agents who already possess market power will not benefit from the control of the enforcement policy and would be less likely to form an SRO. For example, in the property/casualty insurance industry, enforcement is also directed toward the customers, in this case policyholders, who also have the opportunity to commit fraud. But since the insurance industry is competitive, policyholders would not gain any additional rents from lax enforcement and therefore need not seek to influence the enforcement policy. Thus we predict that enforcement policies directed at policyholder fraud are efficient whereas those directed at insurance broker fraud may be inefficient.

One aspect of enforcement that is absent from our analysis concerns the fact that customers may update their beliefs regarding the industry depending on enforcement outcomes. If a broker or dealer is caught cheating, customers may revise upwards the likelihood of being cheated or the likelihood that cheaters are caught. To incorporate the former, one could assume that some fraction of agents are honest (they tell the truth irrespective of the payoffs) and that customers are uncertain of this fraction and update based on history. To incorporate the latter effect, one could assume that the SRO's enforcement policy is not perfectly observed. Both of these factors would affect the SRO's optimal enforcement policy and including them would generate fraud in equilibrium, something also absent from our model.

We assumed that the parties to the contract did not directly choose the enforcement policy. While this does resemble standard practice, it would be interesting to analyze why the enforcement policy is delegated to some other institution, whether it be the state or an SRO. The explanation may simply be that a specialized institution can investigate for fraud at a lower cost than can individual parties to a contract. Beyond cost advantages, though, are there any important externalities across contracting parties that would give rise to this separation between contracting and enforcement? Identifying these externalities and incorporating them in a model of contract enforcement is an important avenue for future research.

7. Appendix

Proof of PROPOSITION 1: This a standard result in mechanism design and we omit the proof. Much of the analysis is contained in Step 1 of the proof of **PROPOSITION 2.**

Proof of PROPOSITION 2: First we establish that customer will always receive at least $\underline{w}$.

LEMMA 1. Suppose (z,r) solves CP(p, x). Then $z(r) \geq \underline{w}$.

Proof of **LEMMA 1**: Define z' such that $z'(r) = \max[z(r), \underline{w}]$. Then note that $v(w,r|z',p,x) \leq v(w,r|z,p,x)$ for all w, r. Define $Q = \{w : z(r(w)) < \underline{w}\}$. Then let

$$r'(w) = r(w) \qquad \text{if } w \notin Q, \text{ and}$$

$$r'(w) = \text{argmax}_r \; v(w,r|z',p,x) \qquad \text{if } w \in Q.$$

Suppose $w \notin Q$ so that $z(r(w)) \geq \underline{w}$. Then, if $z(s) \leq w$,

$$v(w,r'(w)|z',p,x) = v(w,r(w)|z',p,x) = v(w,r(w)|z,p,x) \geq v(w,s|z,p,x) \geq v(w,s|z',p,x).$$

Hence, (z',r') satisfies (AIC) and is feasible for CP(p,x). Finally, $z'(r'(w)) = z'(r(w)) = z(r(w))$ if $w \notin Q$. However, for $w \in Q$, $z'(r'(w)) \geq \underline{w} > z(r(w))$. Thus, $E[\, z'(r'(W))\,] > E[\, z(r(W))\,]$ unless $Q = \varnothing$. Hence if (z,r) solves CP, we must have $Q = \varnothing$. ◆

Now consider the following problem:

SRP′: $\max_{z,r,p,x,t} \quad E[\, v(W,r(W)|z,p,x)\,]$

s.t. (AF), (AIC), (CIR), (RB) and

(ZW*) $z(r) \geq \underline{w}$.

SRP′ has the same objective as SRP but since (CIC) implies (AF) and (AIC) by definition, and (ZW*) by **LEMMA 1**, SRP′ is less constrained than SRP. So if a solution to SRP′ satisfies the constraints of SRP then this is also a solution to SRP. We will first show that SRP′ simplifies to SRP* with $r(w) = w$, $t = E[\, p(W)\, c\,]$, $x(w,w) = 0$ and $x(w,w') = w - z(w')$ for $w' \neq w$. We next show that if (z,r,p,x,t) is a solution to SRP′ then (z,r,p,x^*,t) is a solution to SRP′. Last we show that (z,r,p,x^*,t) satisfies the constraints of SRP.

Step 1. Establish that the solution to SRP′ is equivalent to the solution to SRP* with $r(w) = w$, $t = E[\, p(W)\, c\,]$, $x(w,w) = 0$ and $x(w,w') = w - z(w')$ for $w' \neq w$.

Suppose (z',r',p',x',t') is feasible for SRP′. We first show that there is a feasible (z,r,p,x,t') with the same payoff and $r(w) = w$. Define (z,p,x) via $z(w) = z'(r'(w))$, $x(w,s) = x'(w,r'(s))$ and $p(w) = p'(r'(w))$. Clearly, $z(w) = z'(r'(w))$ implies that (AF), (CIR) and (ZW*) are satisfied. Next, note that

$$\begin{aligned} v(w,s|z,p,x) &= p(s)\, u([w - z(s) - x(w,s)]^+) + (1-p(s))\, u(w - z(s)) \\ &= p'(r'(s))\, u([w - z'(r'(s)) - x'(w,r'(s))]^+) + (1-p'(r'(s)))\, u(w - z'(r'(s))) \\ &= v(w,r'(s)|z',p',x'). \end{aligned}$$

Thus if $z(s) = z'(r'(s)) \leq w$, we have $v(w,s|z,p,x) = v(w,r'(s)|z',p',x') \leq v(w,r'(w)|z',p',x') = v(w,w|z,p,x)$ so that (AIC) is satisfied. Finally, note that

$$p(w)(c - \min[x(w,w),\, w - z(w)]) = p'(r'(w))(c - \min[x'(w,r'(w)),\, w - z'(r'(w))]),$$

so that (RB) is unchanged. Hence, (z,r,p,x,t') is feasible for SRP′ and since $v(w,w|z,p,x) = v(w,r'(w)|z',p',x')$, the agent's payoff is unchanged. Thus we have shown that we can without loss of generality assume $r'(w) = w$. We thus start over with this assumption.

Suppose (z',r',p',x',t') is feasible for SRP′ with $r'(w) = w$. Consider (z,r',p',x,t) where

$$z(w) = z'(w) + p'(w)\, \min[x'(w,w),\, w - z'(w)],$$

$x(w,w) = 0$, $x(w,w') = w - z(w')$ for $w' \neq w$, and $t = E[\, p'(W)\, c\,]$. Note that $x'(w,w) \geq 0$ implies that $z(w) \geq z'(w) \geq \underline{w}$ and (ZW*) holds. Also, $z(w) \leq z'(w) + w - z'(w) = w$ and (AF) holds. Next note that

$$E[\, z(W)\,] - t = E[\, z'(w) - p'(w)(c - \min[x'(w,w),\, w - z'(w)])\,] \geq E[\, z'(w) - t'\,] \geq \alpha$$

and (CIR) holds. Since $r'(w) = w$ and $x(w,w) = 0$, (RB) holds. Since u is concave,

$$\begin{aligned} u(w - z(w)) &= u(w - z'(w) - p'(w)\, \min[x'(w,w),\, w - z'(w)]) \\ &\geq p'(w)\, u([w - z'(w) - x'(w,w)]^+) + (1-p'(w))\, u(w - z'(w)) \\ &= v(w,w|z',p',x'). \end{aligned}$$

Thus, the agent's payoff is higher. It remains to check (AIC). If $z(s) \leq w$ then

$$u(w - z(w)) \geq v(w,w|z',p',x') \geq v(w,s|z',p',x')$$

$$= p'(s)\, u([w - z'(s) - x'(w,s)]^+) + (1-p'(s))\, u(w - z'(s))$$

$$\geq p(s)\, u(0) + (1-p(s))\, u(w - z(s)),$$

and (AIC) follows since $u(0) = 0$. Thus we have shown that we can restrict attention to solutions with $r(w) = w$, $x(w,w) = 0$, $x(w,w') = w - z(w')$ for $w' \neq w$, and $t = E[\, p'(W)\, c\,]$. Substituting this r, x, and t into SRP′ yields SRP*.

Step 2. Establish that if (z,r,p,x,t) with $r(w) = w$, $x(w,w) = 0$, $x(w,w') = w - z(w')$ for $w' \neq w$, and $t = E[\, p'(W)\, c\,]$ is a solution to SRP′ then (z,r,p,x^*,t) is also solution to SRP′.

This follows because with a switch from x to x^*, (i) (AF), (CIR) and (ZW*) are unaffected; (ii) with $r(w) = w$ and $x^*(w,w) = x(w,w) = 0$, the objective and (RB) are unaffected; and (iii) by the definition of x^*, (AIC) holds.

Step 3. It remains to show that (z,r,p,x^*,t) satisfies the constraints of SRP.

Since the difference between SRP′ and SRP is that the former drops the constraint (CIC), we need only show that (z,r,p,x^*,t) satisfies (CIC), or equivalently, (z,r) solves CP(p,x^*).

First note that x^* is well-defined and that for $z(w') < z(w)$, $0 < x^*(w,w') \leq w - z(w')$. Define $\underline{z}$ by $\underline{z}(w) \equiv \underline{w}$. By the definition of x^*, for $z(w') < z(w)$,

$$v(w,w|z,p,x^*) = v(w,w'|z,p,x^*) \leq v(w,w'|\underline{z},p,x^*).$$

Also, for $z(w') \geq z(w)$,

$$v(w,w|z,p,x^*) = u(w - z(w)) \leq u(w - \underline{w}) = v(w,w'|\underline{z},p,x^*).$$

Thus, we have shown that for all w', $v(w,w|z,p,x^*) \leq v(w,w'|\underline{z},p,x^*)$.

Now suppose the customer proposes an alternative (z',r') satisfying (AF) and (AIC) given (p,x^*). By (AF), $z'(r'(w)) \leq w$, so by Lemma 1 we can assume $z'(r'(\underline{w})) = \underline{w}$. Let $\underline{r} = r'(\underline{w})$. Then for $w \geq \underline{w}$,

$$v(w,r'(w)|z',p,x^*) \geq v(w,\underline{r}|z',p,x^*) = v(w,\underline{r}|\underline{z},p,x^*) \geq v(w,w|z,p,x^*)\ ,$$

where we use the previous result for the last inequality. Thus the agent must be better off under (z',r') than under (z,r). We now show that this implies the customer must be worse off. We can rewrite the above as,

$$p(r'(w))\, u(w - z'(r'(w)) - x^*(w,r'(w))) + (1-p(r'(w))\, u(w - z'(r'(w))) \geq u(w - z(w)).$$

Since u is concave, this implies

$$z'(r'(w)) + p(r'(w))\, x^*(w,r'(w)) \leq z(w). \qquad (7)$$

Since $x^* \geq 0$, this implies $z'(r'(w)) \leq z(w)$, and the customer is worse off for all $w \geq \underline{w}$. Thus, (z,r) solves $\mathrm{CP}(p,x^*)$.[20] ◆

Proof of PROPOSITION 3: Suppose $\alpha \leq \underline{w}$. The solution to SRP^* is immediate. The solution to MA is at least as good as the solution $z \equiv \alpha$ and $p \equiv 0$ which is better than the solution to SRP^*.

Now suppose $\alpha > \underline{w}$. Suppose (CIR^*) does not bind. Then define $z'(w) = \min[z(w), b]$ for b such that $E[\, z'(W) - p(W)c\,] = \alpha$; i.e., such that (CIR^*) binds. Clearly (AF^*) and (ZW^*) hold for (z',p). For (AIC^*), if $z'(w') < z'(w)$, then $z'(w') = z(w')$. Hence,

$$u(w - z'(w)) \geq u(w - z(w)) \geq (1-p(w'))\, u(w - z(w')) = (1-p(w'))\, u(w - z'(w')),$$

and (AIC^*) holds. Thus, (z',p) is feasible for SRP^*, and since $z'(W) \leq z(W)$ and $E[\, z'(W)\,] < E[\, z(W)\,]$, it has a higher agent payoff. Therefore, at any solution to SRP^*, (CIR^*) must bind.

The monopolist agent's problem, MA, is equivalent to SRP^* without the constraint (ZW^*). Hence if at a solution to MA, (ZW^*) is not binding, then the solutions to the two problems coincide. Suppose (z, p) satisfies (AF^*), (AIC^*) and (CIR^*). Define $z'(w) = \max[\underline{w}, z(w) - a]$, where the constant a satisfies $E[\, \max[\underline{w}, z(w) - a]\,] = E[\, z(w)\,]$. Note that (CIR^*) and $\alpha \geq \underline{w}$ imply that $a \geq 0$. Since $E[\, z'(w)\,] = E[\, z(w)\,]$, the value of the objective is the same for (z', p) and (z, p); and (z', p) satisfies (CIR^*). We now show that (z', p) also satisfies the other constraints of SRP^*. Since $a \geq 0$, we have $z'(w) \leq \max[\underline{w}, z(w)] \leq w$ and (AF^*) is satisfied. If $z'(w) = \underline{w}$, $w - z'(w) = w - \underline{w} \geq (1 - p(w'))(w - \underline{w}) \geq (1 - p(w'))(w - z'(w'))$ and if

$z'(w) > \underline{w}$, $w - z'(w) = w - z(w) + a \geq (1 - p(w'))(w - z(w')) + a \geq (1 - p(w'))(w - z(w') + a) \geq (1 - p(w'))(w - \max[\underline{w}, z(w') - a]) = (1 - p(w'))(w - z'(w'))$. So (AIC*) is satisfied. Finally, (ZW*) is clearly satisfied. So for any feasible (z, p) for the monopolist agent's problem, there is a (z', p) that is as good and satisfies the constraints of SRP*. Therefore (ZW*) is not binding and the solution to the two problems coincide. ◆

Proof of PROPOSITION 4: Suppose (z,p) is feasible. Then from (AIC*), for $w \geq z(w')$,

$$u(w - z(w)) \geq (1-p(w'))\, u(w - z(w'))$$

which is equivalent to $p(w') \geq 1 - [\, u(w - z(w)) / u(w - z(w'))\,]$. Thus, (AIC*) is equivalent to (after switching w and w'),

$$p(w) \geq \max_{w' \geq z(w)} 1 - [\, u(w' - z(w')) / u(w' - z(w))\,].$$

Note that for $w' = w$, $1 - [\, u(w' - z(w')) / u(w' - z(w))\,] = 0$, and if $w' \geq z(w)$ and $z(w') < z(w)$, then

$$1 - [\, u(w' - z(w')) / u(w' - z(w))\,] < 0.$$

Therefore, (AIC*) is equivalent to

$$p(w) \geq p^*(w) \equiv \max_{z(w') \geq z(w)} 1 - [\, u(w' - z(w')) / u(w' - z(w))\,].$$

Suppose $p(W) > p^*(W)$ with positive probability. Then (z,p^*) is also feasible for SRP*, and since $E[\, p(W)\,] > E[\, p^*(W)\,]$, (CIR*) does not bind for (z,p^*). By **PROPOSITION 3**, this implies that (z,p^*) and hence (z,p) cannot solve SRP*. Finally, $p^*(w)$ decreasing in $z(w)$ is immediate.

By **PROPOSITION 3**, if $\alpha \leq \underline{w}$, then $p = 0$ and does not vary with α or c. Now consider $\alpha > \underline{w}$. With a risk-neutral agent, **PROPOSITION 2** and **PROPOSITION 3** imply that SRP is equivalent to

SRP^{RN}: $\min_{z,p}$ $E[\, z(W)\,]$

subject to (AF*), (CIR*), and

(AIC^{RN}) $w - z(w) \geq (1 - p(w'))\, (w - z(w'))$ for all w' such that $z(w') \leq w$.

---

[20] Inequality (7) also implies that the result does not depend on our assumption that the SRO collects the penalty instead of the customer.

Let (z_i, p_i) be the solution to this problem for $c = c_i$, and suppose $c_2 > c_1$. By **PROPOSITION 3**, (CIR*) binds and thus $E[\, z_2(W)\,] \geq E[\, z_1(W)\,]$.

We now show that $E[\, p_2(W)\,] \geq E[\, p_1(W)\,]$. Suppose instead that $E[\, p_2(W)\,] < E[\, p_1(W)\,]$ and consider a solution (z_3, p_2) for $c = c_1$, where

$$z_3(w) = z_2(w) - E[\, z_2(W) - z_1(W) - (p_1(W) - p_2(W))\, c_1\,].$$

First observe that $E[\, z_3(W)\,] = E[\, z_1(W) - (p_1(W) - p_2(W))\, c_1\,] < E[\, z_1(W)\,]$. Now check feasibility for (z_3, p_2). Since $z_3(w) < z_2(w)$, (AF*) is satisfied. Since $z_3(w)$ equals $z_2(w)$ less a constant, (z_3, p_2) satisfies (AIC*). Since $E[\, z_3(W) - p_2(W)\, c_1\,] = E[\, z_1(W) - p_1(W)\, c_1\,]$, (CIR*) is satisfied. Hence (z_1, p_1) is not a solution for $c = c_1$ and we have a contradiction. An identical argument applies for α. ◆

Proof of PROPOSITION 5: The case $\alpha \leq \underline{w}$ is trivial, so we assume $\alpha > \underline{w}$. If the agent is risk neutral (AIC*) becomes (AIC^{RN}) above, or equivalently,

$$z(w) \leq z^*(w) \equiv \min_{w'} p(w')\, w + (1-p(w'))\, z(w').$$

It is easy to see that z^* is increasing and concave in w and that $w - z^*(w)$ is increasing and convex in w. We will show that at an optimum, $z(W) = z^*(W)$. Suppose instead $z(W) < z^*(W)$ with positive probability. Then we can define z' by $z'(w) = \min[z^*(w), b]$, for b such that $E[\, z'(W)\,] = E[\, z(W)\,]$. Define $p'(w) = p(w)$ if $z'(w) < b$ and $p'(w) = 0$ if $z'(w) = b$. To see that (z',p') satisfies (AIC*), if $z'(w') = b$ then $z'(w) \leq b = p'(w')\, w + (1-p'(w'))\, z'(w')$, whereas if $z'(w') < b$, then $p'(w') = p(w')$ and $z'(w') = z^*(w') \geq z(w')$, so

$$z'(w) \leq z^*(w) \leq p(w')\, w + (1-p(w'))\, z(w') \leq p'(w')\, w + (1-p'(w'))\, z'(w').$$

Given this, it is straightforward to check that (z',p') is feasible for SRP*, and the payoffs of the agent are unchanged. Now replace p' with p'^* defined in **PROPOSITION 4**; since (AIC*) holds for p', then $p'^*(w) \leq p'(w) \leq p(w)$. Moreover, since we have raised z below b and strictly lowered z above b, $p'^*(w) < p(w)$ for $w = \text{argmax}\{z(w') : z(w') \leq b\}$. Hence, (CIR*) does not bind, and by **PROPOSITION 3**, (z',p'^*) and hence (z,p) is not optimal. Thus, at an optimum we must have $z(W) = z^*(W)$. ◆

Proof of PROPOSITION 6: First we formally state the government's mechanism design problem:

$\text{GP}(p_s,x_s,t_s)$: $\max_{z,r,p_g,x_g,t_g} \; E[\, z(r(W))\,] - t_g - t_s$

subject to (CIC′) (z,r) solves $\text{CP}(p_s + p_g,\, x_g)$

(GB) $t_g \geq E[\, p_g(r(W))\,(c_g - \min[\, x_g(W,r(W))\,,\, W - z(r(W))])\,]$

$$-E[p_s(r(W))\max[0,\min[x_g(W,r(W)),W-z(r(W))]-x_s(W,r(W))]]$$

Replace (CIC′) in GP with the constraints (AF) and (AIC). Call this problem GP′. Suppose the government chooses a non-informative reporting strategy, $r(w_1) = r(w_2) = r_1$. Then (AF) and the same argument as in **LEMMA 1** imply that $z(r_1) = w_1$. Thus the agent pays $z_2 = w_1 + p(r_1)\min(x_g(w_2,r_1),\, w_2 - w_1)$ in expectation when $W = w_2$. Since the agent is risk averse, having the agent report $r_2 \neq r_1$, with $p_g(r_2) = 0$, $x_g(w_2,\, r_2) = 0$, and $z(r_2) = z_2$ satisfies (AIC), and raises the customer's payoff (since costly investigations by the government are avoided and $x_s(w_2,\, r_1)$ is not paid to the SRO). Thus we can assume $r(w_1) \neq r(w_2) = r_2$, and by the same reasoning, $p_g(r_2) = 0$, $x_g(w_2,\, r_2) = 0$. Also, $z(r(w_1)) = w_1$, since raising $z(r(w_1))$ until (AF) binds relaxes (AIC) and benefits the customer. Hence, $x_g(w_1,\, r(w_1)) = 0$ without loss of generality. Finally, $x_g(w_2,\, r(w_1)) = w_2$; i.e., the maximum penalty is imposed off equilibrium to relax the (AIC) constraint. Given that, it is clearly optimal for the government to choose $r(w_1) = w_1$. This is because $p_s(w_1) \geq 0 = p_s(r),\, r \neq w_1$; hence the government can get investigations performed by the SRO with probability $p_s(w_1)$ at no direct cost to the government. Thus we can without loss of generality assume $r(W) = W$.

Therefore we have shown that $r(W) = W$, maximum penalties off equilibrium and zero penalties in equilibrium are optimal for the government. Substituting this into GP′ yields GP^*. Finally, note that at this solution (CIC′) holds, and hence GP is equivalent to GP^*. ◆

Proof of PROPOSITION 7: Let $\pi_i = Pr(W{=}w_i)$, $z_i = z(w_i)$ and $p_i = p(w_i)$. From **PROPOSITION 6** and the subsequent discussion in the text, together with the fact that $z \geq w_1$ at a solution to the government's problem (see the proof of **PROPOSITION 6**), GP^* in the binary case reduces to

$\max_{z,r,p}$ $\pi_2\, z_2 + \pi_1\, w_1 - \pi_1\, (p_1 - p_s(w_1))^+\, c_g$

subject to (AIC) $u(w_2 - z_2) \geq (1-p_1)\, u(w_2 - w_1)$.

Here we have used the fact that since $z_2 \geq w_1 = z_1$, only one (AIC) constraint is relevant. Thus, p_2 does not appear in the constraints and it is optimal for the government to choose $p_2 = 0$.

It is immediate that the solution to the above problem has $p_1 \geq p_s(w_1)$. Thus, we can write the objective as

$$\pi_2\, z_2 + \pi_1\, w_1 - \pi_1\, p_1\, c_g + \pi_1\, p_s(w_1)\, c_g.$$

Note that the last term is a constant and can be ignored. Thus, the government's problem is identical to the competitive problem (1) with cost c_g and the constraint $p_1 \geq p_s(w_1)$. Since the problem is concave, its solution is characterized by first-order conditions, so the added constraint either binds or is irrelevant. Thus, the solution to GP is given by

$$p_1 = \max[\, p^{cg}(w_1),\, p_s(w_1)],$$

with z_2 following from (AIC).

Now consider the SRO's problem, SRPO. Since the government will always increase enforcement up to p^{cg}, the SRO will not choose $p_s(w_1) < p^{cg}(w_1)$. This is because $c > c_g$, so that by increasing p_s up to p^{cg} the aggregate transaction tax $t_s + t_g$ can be reduced by $(p^{cg}(w_1) - p_s(w_1))(c_g - c)$. This leaves the agent's payoff per-transaction unchanged, but raises the customer's payoff α and therefore the transaction volume $F(\alpha)$. Hence, $p_s(w_1) \geq p^{cg}(w_1)$. This implies $p_g = 0$; i.e., the SRO pre-empts government enforcement.

Thus, SRPO is equivalent to SRP with heterogeneous agents and the added constraint that $p_s(w_1) \geq p^{cg}(w_1)$. This is equivalent to the constraint $\alpha \geq \alpha^{cg}$, where α^{cg} is the customer's utility when $p_s = p^{cg}$. If $F(\alpha)\, V(\alpha)$ is concave, this constraint is either binding or irrelevant; that is, $p_s = \max[p^{srp}, p^{cg}]$, where p^{srp} is the SRO's unconstrained solution. With $F(\alpha)$ log-concave, concavity follows if $V(\alpha)$ is log-concave. We show next that $V(\alpha)$ is concave, a stronger result. SRP in the binary case is characterized by

$$V = \pi_2\, u(w_2 - z_2), \qquad \alpha = \pi_1\, w_1 + \pi_2\, z_2 - \pi_1\, p_1\, c, \qquad u(w_2 - z_2) = (1-p_1)\, u(w_2 - w_1).$$

Combining these equations and simplifying, it can be shown that

$$V(\alpha) = U(k\,V(\alpha) - \alpha),$$

where U is increasing and concave. Then,

$$V' = U'\,(k\,V' - 1), \qquad V'' = (-\,U'/U'')\,V'\,(k\,V' - 1)^2,$$

and V is concave since $V' < 0$. ◆

Proof of PROPOSITION 8: Since $z_1 = w_1$ by (ZW$^\infty$) and (AF$^\infty$), (CIR$^\infty$) implies $z_2 > w_1$. Hence, (AIC$^\infty$) reduces to a single constraint: an agent with outcome w_2 does not report w_1. Thus, θ_2 can be reduced to zero, raising the objective function and relaxing (AIC$^\infty$). Then p_2 can be reduced to zero, which only relaxes (CIR$^\infty$). But then z_2 can be reduced, raising the objective function. Thus, $p_2 = \theta_2 = 0$. Given that, the (AIC$^\infty$) constraint can be written,

$$w_2 - z_2 + \delta\,V^\infty(\alpha) \geq (1 - p_1)\,(w_2 - w_1 + \delta\,(1 - \theta_1)\,V^\infty(\alpha)).$$

This constraint must bind, since otherwise θ_1 can be reduced to zero, raising the objective function, and then (if it still does not bind) p_1 can be reduced, relaxing (CIR$^\infty$). As above, this allows z_2 to be reduced to raise the objective function, so that (CIR$^\infty$) also binds.

Solving (AIC$^\infty$) for z_2 and plugging this into the objective function then yields the expression for V^∞. Solving for θ_1 given V^∞ and p_1 and plugging into (CIR$^\infty$) leads to the expression for α. ◆

Proof of PROPOSITION 9: Since (CIR$^\infty$) binds, $V^\infty(\alpha)$ must strictly decrease in α. This implies that holding fixed the agent's payoff $V^\infty = V^\infty(\alpha)$, the highest possible payoff for the customer is α. Since (6) is linear in p_1, this implies that the solution occurs at an extreme point of the set of feasible p_1. This feasible set is defined by the restrictions $p_1 \in [0,1]$ and $\theta_1 \in [0,1]$. From (5), if V^∞ is held fixed, θ_1 decreases with p_1. Thus, the possible boundary conditions are

$$p_1 = 1 \text{ or } \theta_1 = 0 \quad \text{if } \pi_2\,[\,w_2 - w_1 + \delta\,V^\infty\,] \geq c,$$

$$p_1 = 0 \text{ or } \theta_1 = 1 \quad \text{if } \pi_2\,[\,w_2 - w_1 + \delta\,V^\infty\,] \leq c.$$

Note that if $p_1 = 1$, θ_1 is irrelevant and can be taken to be zero. Thus the first boundary condition can be restated simply as $\theta_1 = 0$. This occurs if

$$V^{\infty} \geq V^0 \equiv [\, c/\pi_2 - (w_2 - w_1)\,]/\delta,$$

or equivalently if

$$\alpha \leq \alpha^0 \equiv \max\{\, \alpha : V^{\infty}(\alpha) \geq V^0 \,\}.$$

When $V^{\infty} < V^0$, $p_1 = 0$ or $\theta_1 = 1$. The boundary between these occurs at $p_1 = 0$ and $\theta_1 = 1$, or $V^{\infty} = V^1 \equiv \pi_2 (w_2 - w_1)$. Since V^{∞} is decreasing in p_1 and θ_1, we can summarize

$$\theta_1 = 0, \text{ if } V^{\infty} \geq V^0,$$

$$p_1 = 0, \text{ if } V^{\infty} \in [V^1, V^0),$$

$$\theta_1 = 1, \text{ if } V^{\infty} < \min[V^1, V^0].$$

For the comparative static with respect to c, note that V^0 is increasing in c. Also, from (6), $V^{\infty}(\alpha)$ is decreasing in c. Thus, α^0 decreases with c. ◆

Proof of PROPOSITION 10: We can use (5) and the right side (6) to compute the expected payoff for the agent and the customer for a given p_1 and θ_1. Holding p_1 fixed, it is immediate from (5) that V^{∞} decreases with θ_1. From the right side of (6), the customer's expected payoff varies with θ_1 only through the effect on V^{∞}. The derivative of the customer's expected payoff (the right side of (6)) with respect to V^{∞} is

$$-\pi_2 (1 - \delta) + \pi_2 \pi_1 p_1 \delta.$$

Thus, the customer's expected payoff decreases with V^{∞}, and so increases with θ_1 if

$$\pi_1 p_1 \delta < (1 - \delta),$$

which implies the result. ◆

8. References

Ameringer, C. F., State Medical Boards and the Politics of Public Protection. Baltimore: Johns Hopkins University Press, 1999.

Cremer, H., Marchand, M., and Pestieau, P., "Evading, Auditing and Taxing: The Equity-Compliance Tradeoff," Journal of Public Economics, 1990, 43, 67-92.

DeMarzo, P. M. and Fishman, M. J., "Optimal Long-Term Financial Contracting with Privately Observed Cash Flows," Northwestern University and Stanford University working paper, 2000.

Faure-Grimaud, A., Laffont, J.-J., and Martimort, D., "The Endogenous Transaction Costs of Delegated Auditing," European Economic Review, 1999, 43, 1039-1048.

Frankhauser, M. M., Gardner, L. M., McNally, B. F., and Leatherwood, L. A., "Enforcement by Self-Regulatory Organizations: A Growing Impact on Broker-Dealers and their Personnel," in The Securities Enforcement Manual: Tactics and Strategies, ed. by R. M. Phillips. Chicago: American Bar Association, 1997.

Gehrig, T. and Jost, P-J., "Quacks, Lemons, and Self Regulation: A Welfare Analysis," Journal of Regulatory Economics, 1995, 7, 309-325.

Jarrell, G. A., "Change at the Exchange: The Causes and Effects of Deregulation," Journal of Law and Economics, 1984, 27, 273-312.

Kofman, F. and Lawarree, J., "Collusion in Hierarchical Agency," Econometrica, 1993, 61, 629-656.

Leland, H. E., "Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards," Journal of Political Economy, 1979, 87, 1328-1346.

Mahoney, P. G., "The Exchange as Regulator," Virginia Law Review, 1997, 83, 1453-1500.

McCaffrey, D. P. and Hart, D. W., Wall Street Polices Itself: How Securities Firms Manage the Legal Hazards of Competitive Pressures. Oxford: Oxford University Press, 1998.

Mookherjee, D. and Png, I. P. L., "Optimal Auditing, Insurance, and Redistribution," Quarterly Journal of Economics, 1989, 104, 399-415.

Morrison, J. and Wickersham, P., "Physicians Disciplined by a State Medical Board," Journal of the American Medical Association, 1998, 279, 1889-1893.

National Association of Securities Dealers, "Securities Regulation in the United States," NASD, 1996.

Phillips, R. M., "Overview: The Multilayered Securities Enforcement System," in The Securities Enforcement Manual: Tactics and Strategies, ed. by R. M. Phillips. Chicago: American Bar Association, 1997.

Pirrong, S. C., "The Self-Regulation of Commodity Exchanges: The Case of Market Manipulation," Journal of Law and Economics, 38, 141-206.

Saloner, G., "Self-Regulating Commodity Futures Exchanges," in The Industrial Organization of Futures Markets, ed. by R. W. Anderson. Lexington: D. C. Heath, 1984.

Sanchez, I. and Sobel, J., "Hierarchical Design and Enforcement of Income Tax Policies," Journal of Public Economics, 1993, 50, 345-369.

Shaked, A. and Sutton, J., "The Self-Regulating Profession," Review of Economic Studies, 1981, 48, 217-234.

Shuchman, H. L., Self-Regulation in the Professions. Glastonbury: The Futures Group, 1981.

Tirole, J., Hierarchies and Bureaucracies: On the Role of Collusion in Organizations," Journal of Law, Economics and Organization, 1986, 2, 181-214.

Townsend, R., "Optimal Contracts and Competitive Markets with Costly State Verification," Journal of Economic Theory, 1979, 22, 265-293.