Seoul Journal of Business
Volume 18, Number 2 (December 2012)

Honesty and Intermediation: Corporate Cheating, Auditor Involvement and the Implications for Takeoff

BRISHTI GUHA*
Singapore Management University
Singapore

Abstract

We examine honesty and credible auditing in firm-investor relations in a repeated game of imperfect information, embedded in a general equilibrium framework. Informed auditors enhance credibility over a range of audit fees – despite the auditor's incentive to collude – provided the probability of detection is imperfectly correlated across clients.

Auditing can enhance growth especially for a relatively egalitarian distribution of wealth. We show that audit fees must be neither too high nor too low to enhance client credibility, highlight the role of mandatory audit fee disclosure, interpret international differences in shareholding patterns and uncover a possible rationale for audit industry concentration.

Keywords: Corporate governance, auditing, disclosure, inequality and takeoff, general equilibrium, repeated games

INTRODUCTION

Recent corporate scandals have often featured firms cheating their shareholders, with or without auditor involvement. This environment raises issues of trust in firm-investor relations and of "self-enforcing honesty." In particular, low shareholder trust obviously has adverse economy-wide repercussions. Once one expects to be dealt with

* School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903. Email: bguha@smu.edu.sg. I thank two anonymous referees for valuable suggestions.

dishonestly, this distrust discourages investors, leading to a collapse of industry. Honesty and trust, on the other hand, feed off each other. Focusing on the moral hazards faced by firm insiders, we ask when firm insiders behave honestly and characterize possible equilibria in a world with stochastic market outcomes and imperfect information. We also investigate when efficiency can be improved – through market creation or increased output and welfare – by employing informed intermediaries (like auditors) and when the presence of such intermediaries could lead to honest equilibria (taking into account the auditor's possible temptations to collude with clients planning fraud or to extort blackmail from honest ones) for a greater range of parameters than in their absence. We show that this can happen provided that the probability of detecting malfeasance by an auditor is imperfectly correlated across clients, and on a tangent, we provide a moral-hazard related rationale for the concentrated structure of the auditing industry which is empirically observed. We also discuss some policy applications of our model and predictions relating to disclosure of audit contracts, public transparency and the Sarbanes-Oxley reforms. We briefly discuss possible extensions of our work including suggestions on international differences in shareholding patterns. We relate our findings to those of other papers, including empirical papers examining the effect of different legal strength on demand for auditor quality across countries or regions, and those studying the relationship between shareholding patterns and auditor choice.

Our paper differs from most others on corporate governance in that it is also related to development.[1)] The distinction between entrepreneurs, who control firms, and ordinary investors, who do not, turns out to reflect the distribution of wealth. This link leads us into the realm of development economics. In our model the level and distribution of wealth affect the growth of corporations and indeed the very existence of the share market for reasons of credibility, with implications for the processes of industrial takeoff. We focus on the *interplay* of these wealth effects with the institution of credible auditing. In particular, after dealing with how credible auditing can be ensured despite collusion possibilities, we highlight how the emergence of a credible auditor affects industrial takeoff prospects,

1) However there is also a literature on financial markets and economic growth, the most relevant of which we will discuss later in the paper.

entry into entrepreneurship and economic efficiency for any given wealth distribution.

While our basic tool is the infinite-horizon one-sided prisoner's dilemma game, we embed this within a general equilibrium framework. This enables us to establish the existence of a unique equilibrium for any given set of parameters. It also makes it possible for us to endogenize the set of firms that participate actively in the game and the set of investors.

The general equilibrium context of our analysis is crucial because, as the Folk theorem assures us, an infinity of equilibria can be supported in an infinitely repeated game, given a high enough discount factor. The equilibrium we focus on however is the only one that is compatible with balance between demand and supply in a general equilibrium based on economy-wide parameters like the level and distribution of wealth. Moreover, people's beliefs are aligned with the fundamentals of the economy, so that they expect only those equilibria to be chosen that are in line with these fundamentals.

Our paper delivers many interesting results. We find that without credible auditing, too much equality in the wealth distribution can impede economic growth – an effect which works through an interaction of wealth effects with share market imperfections. We then find that (a) audit fees can be neither too low nor too high to enhance client credibility, (b) given mandatory disclosure of audit fees, firms can credibly raise more capital and entry into entrepreneurship increases (c) auditing increases economic growth either if the economy is relatively poor or if there is a sizable middle class (in other situations, auditing may have a largely redistributive effect on payoffs), (d) firms are more eager to hire auditors with a large number of clients when the audit fee regime is transparent, and (e) auditors, especially those with a large number of clients, can serve as "substitute safeguards" for investors and such auditors can coexist along with a widely held shareholding pattern. As we discuss later, there is empirical support for some of these results, while the others are empirically testable. Moreover, our paper is also novel in examining the interaction between wealth distribution and corporate governance/auditing. The finance and growth literature does not do this, as it mainly looks at how various measures of credit market development affect the growth of industries and countries.

The rest of the paper is organized as follows. In section 2 we lay

out our assumptions. Section 3 presents a possible model of insider-outsider interaction in the absence of an external auditor, where the set of insiders is endogenized, and different types of equilibria are characterized. Section 4 discusses some features and developmental implications of this model. In section 5 we introduce auditing and derive conditions for an equilibrium with credible auditing and efficiency gains : we also discuss the sources of the efficiency gains that auditing can achieve, despite possibilities of auditor-client collusion. While all these sections take capital structure as given, concentrating on equity financing, in section 6 we find the optimal capital structure, to which all our previous results remain applicable. In section 7 we propose some extensions dealing with international differences in share financing patterns.

ASSUMPTIONS

We consider a community of infinitely-lived individuals. The wealth of each individual is inelastic and fixed: they can neither save[2)] nor borrow and their capital does not depreciate. However, they can lend (to the outside world or to government) at a fixed interest rate R. All agents are risk-neutral.

Individuals can become entrepreneurs and set up firms, which however each require a minimum investment of $\underline{I}$. Entrepreneurs have no access to any external source of funds other than investors. Problems of collective action and contract enforcement are sufficient to prevent them forming partnerships among themselves. Groups sufficiently cohesive to solve these problems (such as the members of the Zaibatsu, the Chaebol, the mercantile families of India and of the Chinese Diaspora) can be viewed as collective entities with a combined wealth that is the relevant factor in this case. Let F denote the personal funds of individuals or of groups of this kind. Those with $F < \underline{I}$ must therefore go public in order to set up an enterprise.

Each enterprise lasts one period only. Thereafter investors can recover their capital in its original form. A fresh enterprise requires refinancing by investors.

Firms enjoy good luck and earn a rate of profit of G on total capital with probability p, and suffer bad luck with a rate of profit

2) A justification is provided for this assumption in the Appendix

of B otherwise (G > B). H = pG + (1 – p)B > R is the firm's profit expectation on total capital. G and B are exogenous parameters – as for example if output is subject to exogenous shocks while prices are fixed in the world market as in a small open economy.

Investors expect to receive the market rate of return D (dividend) on their capital. The entrepreneur in addition should receive an "autonomous" expected payment of A, which we also endogenize. For the bulk of this paper we shall assume that these expectations are to be met by promising outside investors ("shareholders" who contribute S to the firm's capital) an amount DSG/H when luck is good and DSB/H when it is bad. The firm is to retain (A + DF)G/H and (A + DF)B/H under good and bad luck respectively. Therefore both outsiders' and insiders' payoffs are state-dependent. The outside investor can then expect an income of

$$pDSG/H + (1 - p)DSB/H = DS[pG/H + (1 - p)B/H] = DS$$

if the firm acts as promised while the insider's income expectations amount to A + DF. These sum of these income expectations A + D(F + S) must equal the total expected profits of the enterprise H(F + S), implying A = (H – D)(F + S). This is a pure equity contract.

However, we later consider other contracts that promise the same return to the investor and endogenize the firm's choice between these alternatives.

We assume that B < R. As we will show later, this assumption will be sufficient to ensure that DB/H < R, so that outsiders prefer not to enter the industry if they expect to get only their bad luck dues. (We will argue D can never exceed H so as not to violate the insiders' participation constraint.)

The parameters F, S, G, B and p are publicly known. But the actual fortunes of the firm cannot be observed by outsiders or legally verified. Therefore, the promise is not a fully enforceable contract. The firm can cheat by restricting outside investors to DSB/H even when luck is good.

There is a one-period-lagged publicly observable signal that detects cheating by insiders with an accuracy (probability) of q. All cheating is collectively remembered.

Of our assumptions, the one that needs further discussion is that of no savings. We assume zero savings so as to focus on the distinctive consequences of cheating without our results being

obscured by the changing dynamics of the accumulation and distribution of wealth. A justification for the assumption is provided in the Appendix.

We will need additional assumptions about auditing, but we defer a discussion of these for the present.

THE MODEL WITHOUT AUDITING

A cheating firm immediately gains

$$G(F + S) - (A + DF)G/H - DSB/H$$
$$= (G - B)DS/H.$$

But it risks detection by the public signal. Exposure of firms as cheats compels them to withdraw their capital (without cost) and reinvent themselves as outside investors in other firms.[3] Now the other option firms have is simply to get the outside opportunity cost on their funds. They will take whichever option gives them more. Recalling that q is the probability of being caught by the public signal, and δ is the discount factor, cheating is deterred if

$$\frac{(G - B)DS}{H} \leq \frac{q\delta\{A + DF - \max\{D, R\}F\}}{(1 - \delta)} \quad (1)$$

Let L = G – B.

Assume first that D > R. This implies

$$A \geq \frac{LDS(1 - \delta)}{q\delta H}.$$

Substituting for A in terms of D, one derives the inequality

$$D \leq \frac{H^2(F + S)q\delta}{(F + S)q\delta H + LS(1 - \delta)} = D^*. \quad (2)$$

3) We have an alternative specification in which cheating insiders' wealth can be seized so that they get zero in all future periods. Qualitative differences from the model presented here are negligible, we omit this model here in the interests of space constraints.

Alternatively, we can express this as a ceiling on s, where s stands for S/F, the ratio of outsiders to insider capital:

$$s \leq \frac{q\delta H(H-D)}{[DL(1-\delta)\} - q\delta H(H-D)]} = s^*(D), \tag{3}$$

or as a lower limit to q:

$$q \geq \frac{s(1-\delta)LD}{\delta H(H-D)(1+s)} = q^*(D). \tag{4}$$

(3) or equivalently, (4) is a credibility constraint. If s exceeds the limit set by (3), investors will expect the firm to cheat and will not therefore invest – and the firm knows this. The insider's participation constraint is A ≥ 0: without this, the insider would invest in other firms rather than go into business himself since his income as an outside investor would exceed his income as an insider. A ≥ 0 implies

$$H \geq D. \tag{5}$$

However, as long as D < H, the insider's expected gains will be an increasing function of outside investment: so the profit maximizing firm will expand up to the limit represented by (3), converting this inequality into an equation[4]. s*(D) can then be represented in the positive quadrant of the (s, D) space by a curve that intercepts the vertical D-axis at D = H and declines monotonically to a horizontal asymptote at $D = (q\delta H^2)/\{q\delta H + L(1-\delta)\}$. Intuitively, the moral hazard of the insider rises with D as well as with s. Any increase in D must therefore be offset by a decrease in s if the firm is to be deterred from cheating.

If however this asymptote is less than R, we must redo our

4) The ceiling s* is determined by the parameters of the game, for a given D (which is endogenized in general equilibrium). We rule out equilibria which are not based on fundamentals – such as those in which every one shares a common belief about some other value of s*, not necessarily based on fundamentals, and invests accordingly because of the conviction that every one else shares the same belief. This is easily justifiable if we assume the absence of co-ordination devices: in that case common knowledge of every one else's beliefs is ruled out, so each individual bases his or her behavior on fundamentals.

calculations from the point at which the s*(D) curve dips below D = R. Assuming D < R, inequality (1) now yields

$$D \leq \frac{q\delta H[(F+S)H - RF]}{q\delta HS + LS(1-\delta)}$$

$$\text{Or } q < \frac{RLS(1-\delta)}{\delta H(H-R)(F+S)}$$

$$\text{Or } s \leq \frac{q\delta H(H-R)}{[RL(1-\delta)\} - q\delta H(H-R)]} = s^*.$$

These limits are independent of D, and therefore constant at the values reached by inserting D = R in (3) and (4).

D (or s) also determines the minimum wealth requirement for an entrepreneur to start business Since the firm needs to reach a minimum size $\underline{I}$ to function, it cannot exist unless

$$\frac{\underline{I}-F}{F} \leq s^*. \tag{6}$$

This implies

$$F \geq \frac{\underline{I}}{1+s^*} = F^*. \tag{7}$$

F* is the minimum wealth needed for entry and depends on the level of s (or D).

Let K be the aggregate wealth of the economy and P(W) the fraction of K owned by those with wealth below W. Then the total demand for outside capital generated by the entrepreneurs who can enter is

$$X_d = K[1 - P\{\frac{\underline{I}}{1+s^*(D)}\}]s^*(D) \tag{8}$$

The term in square brackets is the ratio of entrepreneurial capital to the total wealth of the economy. The RHS, therefore, represents the amount of outside capital that entrepreneurs can apply for

without compromising their credibility. As D falls, not only can each firm credibly invite more outside investment, but also more firms can enter and create additional demand for capital.

The total supply of outside capital is the total wealth of those below the threshold for entry:

$$X_s = KP[\frac{\underline{I}}{1 + s^*(D)}]. \tag{9}$$

This is subject to the outside investor's participation constraint D ≥ R which enables them to recover their opportunity cost.

With the supply and demand functions for outsider capital thus defined, two kinds of equilibrium are possible:

First, (case 1) a regular interior equilibrium with H > D > R (figure 1) in which

$$s^*(D) = \frac{P(\frac{\underline{I}}{1 + s^*(D)})}{1 - P(\frac{\underline{I}}{1 + s^*(D)})}. \tag{10}$$

A feature of this is that, for given $\underline{I}$, the equilibrium value of s* is uniquely determined by the distribution of wealth alone. Write P as a function of s*, P = Q(s*), equation (10) then reduces to

$$s^* = Q(s^*)/\{1 - Q(s^*)\}, \tag{10a}$$

which has a unique solution. In this equilibrium, outside capital is fully employed without being in excess demand. The optimal ratio of outsider to entrepreneurial capital exactly corresponds to the ratio of wealth owned by those below the minimum entry requirement to that owned by those above. The interior equilibrium occurs for parameter ranges such that

$$P(\frac{\underline{I}}{1 + s^*(R)}) < s^*(R)[1 - P(\frac{\underline{I}}{1 + s^*(R)})].$$

Second, (case 2) an equilibrium in which D = R, the participation constraint of the investor binds and investors are indifferent

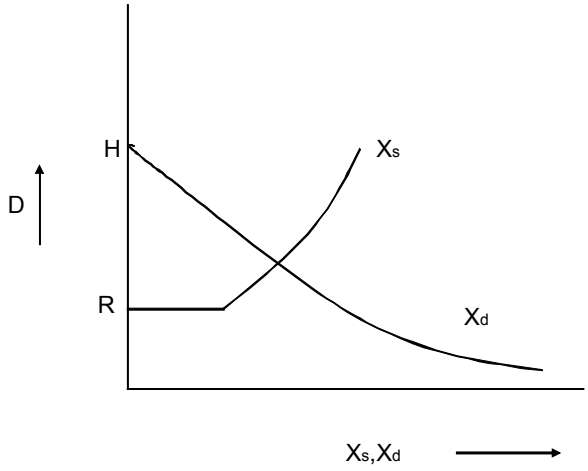

Figure 1.

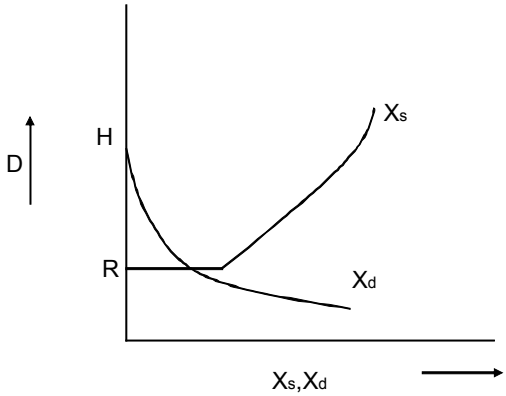

Figure 2.

between investing in the firm and in their outside option[5] (figure 2). Here there is an "excess supply" of capital which has taken shelter in its outside option. This equilibrium occurs for parameter ranges such that

$$P(\frac{I}{1+s^*(R)}) > s^*(R)[1-P(\frac{I}{1+s^*(R)})].$$

5) However, investors will take care that their investment in the firm is not so much that it tempts cheating.

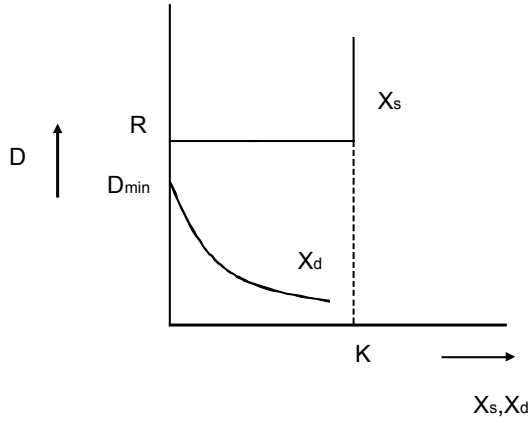

Figure 3.

There also remains a possibility (case 3) that no market may exist. If W_{max} is the wealth of wealthiest individual, $P(W_{max}) = 1$. Suppose that $s^*(D_{min}) = (\underline{I}/W_{max}) - 1$. Then for all $D \geq D_{min}$, $X_d = 0$. The demand curve lies in the positive quadrant only for $D < D_{min}$. Now, if $D_{min} < R$, the demand and supply curves will not intersect. No equilibrium will be possible (figure 3). A low W_{max} implies a higher $s^*(D_{min})$, and therefore a low D_{min}. Thus if the society is a poor one without any rich individuals, a market may not exist.

To sum up, without auditing, only those with wealth above a threshold can become entrepreneurs. Moreover, a market may not exist if no one in the society has wealth above a particular floor. However, if a market exists, there can be two different types of equilibria depending on the distribution of wealth. Only one of these is an interior equilibrium (with insiders earning strictly more than outsiders, who in turn invest all their wealth in the enterprise and earn strictly more than R). In the second equilibrium, outsiders are indifferent between financing the enterprise and the outside opportunity – but some outsider capital is necessarily not invested in the enterprise in the interests of credibility.

SOME FEATURES OF THE MODEL

Before going on to our main results, which concern auditing, we highlight some features of the model without auditing. The model

so far has two sets of features. The first set, related to the idea that a high proportion of outsider financing intensifies moral hazard, is empirically supported by (Joh 2003) and (Lemmon and Lins 2003), who use evidence from Korea to show that in a large sample of externally audited firms, misappropriation by controlling insiders was severe wherever these insiders had a low ownership stake.

Though our model deals with the share market, it is closely connected to theoretical models on imperfect markets for capital or credit. For example, in (Banerjee and Newman 1993), the initial wealth distribution in the population determines occupational choice - only those whose wealth exceeds a certain floor can become entrepreneurs. The underlying causes are imperfect capital markets (as entrepreneurs may renege on loans) and a minimum size requirement for making the enterprise operational. In our model we combine ideas of indivisibilities in enterprise size and imperfect capital markets with the moral hazard that entrepreneurs (insiders in our model) face with regard to their outside shareholders.

This ties up with the second set of features of this model, those related to development. The model illuminates a key impediment to the industrialization of poor countries. In these countries, even when aggregate wealth is adequate for large-scale industry, it could be spread too thinly for effective mobilization. Too few individuals may be wealthy enough to either enter large-scale industry on their own or to attract enough outside funds to do so. The limits to borrowing are well-known. We show here the existence of a constraint on raising capital from the share market that operates through a limit on the ratio of outsider to insider capital set by concerns over cheating. This is one of the main causes of the thinness of share markets in most poor countries,[6] It accounts for the dominance of extended families in the early industrialization of such countries, the role for example of business dynasties in nineteenth and early twentieth century Japan and twentieth century Korea, India and the Chinese business sphere, since capital flowed freely within these families because of a degree of trust and reciprocity among the members. (Burkart, Panunzi and Shleifer 2003) have a somewhat related explanation of the origin of family

6) Empirical work links the extent to which firms go public, as opposed to operating primarily as family-owned firms, to the extent of shareholder protection available. This becomes relevant to our work as we show later that credible external auditing can create a market and facilitate raising capital in the share market.

firms; in their story, family firms are a second best solution in environments where poor investor protection limits the founding family's ability to control expropriation by a manager. Our model, in contrast to theirs, emphasizes the role of indivisibilities and of the distribution of wealth. [Other theoretical explanations of family firms exist, however, some of which are surveyed in (Bertrand and Schoar 2006). For example, if talent is inherited, and if founding a firm requires talent, family firms may have access to a better talent pool. Moreover, family firms might take a more long-run perspective to management if they are concerned about their heirs. Some explanations centre on kinships between business and politics in countries where the latter has an important impact on the former (Faccio 2006)].

Our model also explains the strategy of governments like the Korean to deliberately foster inequality so as to facilitate the accumulation of personal fortunes that could help in building up credible large-scale industries.[7)]

If, however, a market exists, and particularly if we have an interior equilibrium of the kind depicted in figure 1, the role of wealth inequality changes. Here, the equilibrium ratio of outsider to insider capital s* is, for any given $\underline{I}$, uniquely determined by the distribution of wealth; and it can be shown that the more egalitarian the distribution of a given aggregate wealth (in the sense of a higher P( ) for any W), the higher must s* be in equilibrium: the demand curve for outside capital $X_d(D)$ will lie further to the left, the supply curve $X_s(D)$ further to the right, so that D will be lower. Indeed, as the distribution of wealth becomes more equitable, the equilibrium level of D may fall to R (as in figure 2). This leads to inefficiency as some wealth is then necessarily invested in the outside option, earning R instead of being invested in industry, earning a higher rate of return, H. Of course, any increase in equity beyond this point leads to the disappearance of the market.

Initial wealth inequality therefore has a positive effect on overall income for two reasons, *up to* a certain level of inequality at which "excess supply" of outsider capital is eliminated. The positive effect at small levels of inequality stems from (a) the necessity of some individuals owning enough personal wealth to meet the entrepreneurial floor, which is necessary for industry to take off

7) (Lal and Myint 1996) provide a good discussion of this.

and (b) even beyond this, an increase in inequality can ensure that an interior equilibrium obtains rather than an excess supply equilibrium – tantamount to an increase in efficiency as it implies all wealth would then be invested in industry rather than some being consigned to the less profitable outside option. However, once inequality becomes large enough to move the economy into an interior equilibrium, further increases in wealth inequality have *no* effect on overall economic efficiency, merely having a redistributive effect through the impact on D. To focus on credibility issues, we have not made complicating technological assumptions apart from assuming positive minimum size requirements on enterprise. Had we assumed increasing or decreasing returns to scale, for example, it is likely that inequality would have had further effects beyond the point at which the supply of outsider capital can fully be absorbed into industry – but our simple technology of constant returns to capital, our only factor, precludes these possibilities.

This view of equity as a barrier to industrialization – at least at low levels of inequality – contrasts starkly with the received wisdom articulated, for example, by (Murphy, Shleifer and Vishny 1989). They see equity as laying the foundations of industrialization by creating a large homogeneous mass market for manufactures. This, however, is only a demand phenomenon and can influence production only under autarchy, since here consumption and production patterns must coincide. In a small open economy, the two are independent.

Empirical evidence on the effect of initial inequality on growth is mixed. While earlier cross-sectional studies tended to suggest a negative relationship between inequality and growth, (for example, (Persson and Tabellini 1994)), *other* work seems to indicate otherwise. (Forbes 2000) and (Li and Zou 1998) discover a positive relationship. They use fixed effects and trace the negative relationship in earlier studies essentially to omitted variables. (Deininger and Squire 1998) and (Barro 2000) find mixed results for panels while (Banerjee and Duflo 2003) find that inequality as such is neutral in its impact on growth, though changes, both positive and negative, in inequality tend to erode growth. While there have been many traditional theoretical arguments in favor of a negative impact of inequality on growth,[8] the theoretical literature has also

8) Apart from the Murphy et al view, which as we have mentioned applies to a

been mixed. (Aghion and Williamson 1998) suggested that inequality favors industrialization in the presence of start-up costs, also a feature of our model. Unlike Aghion and Williamson, we abstract from decreasing returns to scale to focus on the moral hazard aspects of the problem, which is why we find that inequality stops having any aggregate effect when all wealth becomes invested in industry. Also, in our model the effect of small levels of inequality operates not just through the ability to start the enterprise but also by making it unnecessary for outsiders to keep some wealth in the less profitable option. Both these factors may serve to offset potential negative effects of inequality, resulting in mixed empirical evidence.[9)10)]

A wealthier economy with the same degree of inequality in initial asset distribution – i. e. with more total wealth but the same Lorenz curve – would have the opposite features. The proportion of total wealth P(W) owned by people with less than a given level W of personal wealth will be lower; given $\underline{I}$, outsider capital requirements will be lower in equilibrium, thus permitting a higher expected income D for outside investors. It becomes easier to sustain a credible capital market – an addition to the long list of factors that tend to make industrialization a cumulative process.

Finally, an increase in minimum firm-size requirements $\underline{I}$ with the same level and distribution of wealth will increase P for any given s*, thereby depressing the demand curve for outside capital and driving up the supply curve, reducing D and increasing s* in equilibrium. Thus, technological indivisibilities make it more likely that markets would collapse.

Our model also relates to, but is different from, the finance and growth literature that links financial market development with economic growth. As an example consider (Rajan and Zingales 1998). They show empirically that industries dependent on external finance grow quickly when financial markets are more developed –

---

closed economy, there are political economy arguments that inequality leads to redistributive policies which hamper growth (variants of which are presented in (Alesina and Rodrik 1994) and (Persson and Tabellini 1994)) – though (Benabou 2000) has argued that neither of these premises holds true in the data.

9) Admittedly, our problem is considerably simplified because of the static nature of the wealth distribution.

10) In our world, a policy-maker who lexicographically prefers income maximization to equity, would choose an optimal wealth distribution – one with just enough inequality to eliminate an excess supply equilibrium in favor of an interior one.

as measured by the ratio of the sum of stock market capitalization and domestic credit to GDP. However our model endogenizes the development of the share market in terms of parameters like the wealth distribution and industry startup costs. Interestingly, however, Rajan and Zingales also show empirically that better accounting standards have the same positive effect on growth. This is consistent with our model; if we interpret better accounting standards as a higher q, s* increases for any given D, shifting up the demand curve for capital and making it more likely that (a) an equilibrium exists, so that industry can take off, and (b) the equilibrium is an interior equilibrium so that all capital is invested in industry instead of lying idle in the outside option. This would boost economic growth.

THE ROLE OF THE AUDITOR

The induction of an auditor who can detect and expose cheating by firms changes the picture. We now assume that *auditing expertise* (an advantage in investigating firms who report having had bad luck and determining if they are cheating) is exogenously distributed in the population as a binary variable taking on the values of 0 or 1. Individuals with 0 auditing expertise can never become auditors, while those with an expertise of 1 can. To simplify matters we assume that the outside option on their auditing time is zero. Each auditor inelastically offers his services to firms – one auditor can service many firms.

In what follows we pinpoint policies which *guarantee* that auditing has beneficial effects on honesty and efficiency and remains credible *in spite of* the possibility of auditor-client collusion. We show that credible auditing facilitates Pareto improvements in equilibrium by relaxing the credibility constraint. s*(D) increases for any level of D. Firms can mobilize more outside capital for a given rate of payout. Also, the minimum wealth level required for entry falls, so that more entrepreneurs enter. We prove this below and then discuss the consequences for the different kinds of equilibrium specified above.

Let V denote the fee to be paid at the outset of each period to a firm-hired auditor. For the time being, we defer a discussion of how many auditors are actually in business. If the investors receive low returns, the auditor investigates and then delivers an audit report

on whether the firm cheated or has just been unlucky. At this stage, the firm and the auditor can bargain with each other, the firm offering a bribe in exchange for a favorable report, the auditor demanding extra payment for such a report. We discuss below the feasible strategies open to each of the actors in this scenario.

First, *firms* may or may not hire auditors at time t = 0, paying an agreed fee. In either event, they may or may not cheat. If they have hired an auditor and cheated, they may choose to bribe him by offering extra payment concurrently with the delivery of a favourable audit report, or they may not. They could make the offer right at the outset (at the time of hiring) or later after the cheating and perhaps the investigation has occurred. If they have hired an auditor and not cheated, and the auditor demands an extra payment for certifying to the fact, they may pay or may not.

Second, the *auditor* accepts a fee at t = 0 and checks whether cheating has occurred. If it has, he could truthfully report the fact or suppress it for a bribe. If it has not, he could truthfully report this without additional demands or demand a payment for such a report. He also however has the option of negotiating at the outset of the period with the firm, offering it a clean report card in exchange for a bribe, both to be delivered at the end of the period.

Third, *investors* observe at t = 0 whether firms have hired auditors or not. Depending on the information regime assumed, they may or may not get to know the auditor's fees. They then decide whether to invest in a firm or not. At t = 1, they may reinvest – or they may not. The information they have at this moment includes the return they have received last period, the ratio of insider to outsider capital, the auditor's report and the public signal.

Proposition 1: With auditing, there exists an equilibrium where (a) all firms engage auditors and act honestly, (b) the auditor neither colludes with cheats, nor does he extort by threatening to blacklist honest firms, (c) investors know this and finance the industry, (d) the investors' off-equilibrium strategy is to shun any firm which is not the auditor's client, to withdraw from any firm the auditor labels a cheat, and to mistrust the auditor if and only if he is revealed by the public signal to be colluding with or blackmailing a client. If audit fees lie in a certain range, and this is disclosed to investors, the honest equilibrium is assured.

For given D, credible auditing (1) raises the ceiling on s below which investors will be able to finance the industry without being

cheated and (2) lowers the floor on the entry requirements for entrepreneurship – even if collusion and extortion are strategies open to the auditor..

Proof: Becoming an auditor's client and staying honest thereafter is more attractive for firms than not hiring an auditor if and only if

$$A + \{D - \max(D, R)\}F - V > 0 \tag{11}$$

or substituting for A, and using D ≥ R in equilibrium,

$$(H - D)(F + S) - V > 0. \tag{12}$$

If firms are to have no incentive to cheat after becoming the auditor's client (and risking certain exposure by the auditor), we require:

$$\frac{[(H - D)(F + S) - V]\delta}{1\text{-}\delta} > \frac{DSL}{H}. \tag{13}$$

If V were 0, the LHS would be the present value of future honesty which we denote by P_h, so that (13) boils down to

$$P_h - \frac{\delta V}{1 - \delta} > \frac{DSL}{H}. \tag{13a}$$

This caps the audit fee that could sustain an honest equilibrium. The auditor's fee however must not only be positive but also sufficient to deter collusion or extortion by him if he is to be credible.

Colluding firms and auditors are vulnerable to exposure by the public signal. The firm takes this into account: the maximum bribe it is willing to pay the auditor is its cheating gain less expected loss due to possible exposure plus expected saving in that event of future audit fees. The auditor compares this bribe with his possible loss of future fees from all his N clients[11] who will dismiss him, once exposure undermines shareholder confidence in him. In the circumstances, the auditor may ask for bribes, not from one, but from all his clients in exchange for collusion with all [we show in the appendix that this is generally the optimal course for an auditor who proposes to collude; however our results are not dependent on

11) 10. We will shortly discuss how N is determined.

multiple collusion, as we explain in a later sub section.]

Collusion offers must be made in advance (so that firms can cheat if they so desire) – but implemented only at the end of the period by simultaneous exchange of bribes for good reports.

Suppose then that before firms announce their payouts the auditor can make a secret offer of a favourable report to each his clients in exchange for a bribe to be paid synchronously with the delivery of the report. If this were possible, the auditor might be able to extract bribes from all his clients (rather than from just one) and would stand to lose fees from all if caught. Each client of course decides independently on the auditor's offer.

How does q, the probability of detection of collusion, change as the number of collusions increases? Detection in a single case destroys the credibility of the auditor and his relationship with all his clients. Assume that the auditor's probability of being caught is imperfectly correlated across firms: q for the auditor is an increasing function of N*, q(N*), where N* is the number of clients who decide to collude with him. We defer for the moment the question of how N* is determined. An extreme example is a scenario in which the probability of not being caught while colluding with any one firm is independent of the probability of not being caught while colluding with any other. Here, we would have

$$q(N^*) = 1 - \{1 - q(1)\}^{N^*}.$$

The most that each firm can offer as a bribe equals gains from cheating less expected loss if caught (the discounted value of future payoffs the entrepreneur could have got as an insider minus the fees he would have had to pay the auditor if both were still in business): the relevant level of public signal accuracy here is q(1), the signal that guides investors. Thus the no collusion constraint is:

$$\frac{q(N^*)\delta V}{1-\delta} > \frac{DSL}{H} - q(1)P_h + \frac{q(1)\delta V}{1-\delta}$$

$$\text{or} \quad \frac{[q(N^*) - q(1)]\delta V}{1-\delta} > \frac{DSL}{H} - q(1)P_h. \qquad (14)$$

(13a) and (14) together impose the following range of inequalities on auditor fees:

$$P_h - \frac{DSL}{H} > \frac{\delta V}{1-\delta} > \frac{\frac{DSL}{H} - q(1)P_h}{q(N^*) - q(1)}. \tag{15}$$

This range is non-empty if and only if

$$q(N^*)P_h > [1 + q(N^*) - q(1)]\frac{DSL}{H}. \tag{16}$$

If q(N*) = q(1), this last inequality implies that firms would have no incentive to cheat even without an auditor. If q(N*) > q(1), firms might cheat in the absence of an auditor, but not under the eyes of one who receives a fee in the appropriate range.

The implied upper limit on the ratio of outsider to insider capital is

$$s^m = \frac{q(N^*)\delta H(H-D)}{DL(1-\delta)[1 + q(N^*) - q(1)] - q(N^*)\delta H(H-D)}. \tag{17}$$

Now

$$s^m > s^* = \frac{q(1)\delta H(H-D)}{DL(1-\delta) - q(1)\delta H(H-D)}$$

if and only if

$$q(N^*)[DL(1-\delta) - q(1)\delta H(H-D)]$$
$$> q(1)[DL(1-\delta)[1 + q(N^*) - q(1)] - q(N^*)\delta H(H-D)]$$

or if and only if

$$(1 - q(1))(q(N^*) - q(1))(1-\delta)DL > 0 \tag{18}$$

Thus the auditor raises the limit on outsider financing compatible with honesty – and lowers the floor on wealth required to become an entrepreneur – provided the public signal is imperfect and his probability of being caught is an increasing function of the number of clients he attempts to collude with.

If collusion can be prevented, auditing has a larger role the noisier the public signal (the smaller is q(1)). Control of collusion,

however, is facilitated by fees in the appropriate range, by a more accurate public signal and by greater patience on the firm's part: s_m, the credible limit on the ratio of outsider to insider capital, is an increasing function of q(1) and q(N*) and of δ.

We have yet to consider the possibility of extortion by auditors. Auditors may attempt extortion from an honest but unlucky firm by threatening to falsely report that it had cheated. However, the firm being blackmailed would recognize the emptiness of this threat. It realizes that if it refuses to pay, the auditor has no incentive to actually implement its threat: while the auditor does not gain anything from lying about the firm (given the latter's refusal to pay), he stands to lose his reputation – and therefore his future clientele – if his lying is exposed by the public signal. Thus, in a subgame-perfect equilibrium extortion is ruled out.

Q.E.D.

Turn now to the determination of N* and N. Firms decide independently on the collusion proposal and unanimity is not guaranteed. However, they differ only in size. Moreover, all firms have an incentive to drive s to the common credibility limit determined by the market parameter D. Thus, ultimately, firms differ only in the volume of entrepreneurial capital F and all differences in their behavior should be traceable to differences in F. Now, every expression involving capital in all our behavioral inequalities is linear homogeneous in F and S – therefore in F (since s = S/F is the same for all firms). Accordingly, any proposal that makes auditing fees and bribes proportional to insider capital and is acceptable to one firm will be acceptable to all. Therefore N* = N – the entire clientele of the audit firm.

We now come to the long deferred question of how N is determined – how many auditors are actually hired and what is the clientele size of each? In the beginning of this section, we have already spelled out our assumptions on the supply of auditing expertise. Suppose M individuals in the population have auditing expertise of value 1, and each inelastically offers his services to firms. Now firms know that an auditor can service several clients, and that each auditor must be paid a credibility wage (corresponding to the Shapiro-Stiglitz efficiency wage) of at least $\underline{V} = V = \{(1 - \delta)/\delta\}[\{(DSL/H) - q(1)P_h\}/\{q(N) - q(1)\}]$ [this is derived from (15), substituting N for N*]. Recall that this credibility wage is the minimum wage auditors must be

paid to guarantee that they have no incentive to collude, and thus to make auditing credible. Now as in all models with an efficiency wage flavor, there is some unemployment, that is a large number of individuals with auditing expertise will not actually be hired. In this situation of excess supply of auditing expertise, the auditors' wage will go down to the lower limit of $\underline{V}$ although it cannot go any lower due to reasons of credibility. Examining the expression for $\underline{V}$, we see immediately that it is decreasing in q(N) and hence in N : firms also realize this and figure out that the credibility wage they have to pay auditors will be lower if they can hire an auditor with a large number of other clients. The intuition behind this is the following : if an auditor has a large number of clients his costs from colluding increase sharply so that even if his wage is not very high, his expected losses from collusion become high enough (due to high probability of getting caught) to deter collusion. So an auditor with more clients needs to be paid a relatively small credibility wage.[12] Due to this, each firm seeks to hire an auditor with as many clients as possible. If there were no limit on how many clients a single auditor could service, what would in fact emerge would be a monopoly in the auditing industry. However more realistically suppose that a single auditor faces an upper limit L on the number of clients it can service: in this case what would emerge is an *oligopoly* where firms each hire just enough auditors such that N = L (each auditor services the maximum number of clients he is technologically capable of servicing) and the number of auditors hired is equal to the total number of firms divided by L; the rest remain unemployed. This concentrated auditing structure that emerges in our model is consistent with the fact that auditing is in reality a highly concentrated industry. In the US, for example, the Big Four auditing firms (Deloitte, Ernst & Young, KPMG and PriceWaterhouse Coopers) audited 99% of all public company sales in 2003 (Cunningham 2006). The same paper mentions that the Herfendahl – Hirshman index measuring market concentration was well above 1800 for auditing in 2006, indicating a very highly concentrated market structure. In many other countries too auditing was controlled largely by a small group of four to six big firms. We do not of course deny that there may be other reasons as well for this high degree of concentration.

12) Please see the next paragraph for a discussion of this point.

Our finding may at first seem at odds with the observation that the biggest auditing firms generally charge higher fees than other auditors. However, on closer inspection we find that the credibility wage $\underline{V}$ is in fact linear homogeneous in S and therefore in F, the entrepreneur's personal wealth. Therefore, though the credibility wage that a particular entrepreneur must pay is lower if he hires an auditor with many clients (relative to if he hires one with few), we could observe a cross-sectional pattern of large auditors being paid higher fees if relatively rich entrepreneurs (that is, those with high F) hired the largest auditors. Thus this pattern does not in fact contradict our model. In addition, there may be other factors, such as differences in the quality of auditing expertise, that may also generate differences in fees across large and small auditors. Our model does not deal with such factors as we assume homogeneous audit quality.

Auditing is depicted by us as a commitment device and audit fees are the cost of credible commitment. This is why auditors have to be paid a credibility wage of $\underline{V}$ which is strictly higher than the outside option for their services. Another implication of this is that the auditing structure which emerges is highly concentrated, mirroring reality.

This implication, however, only holds when firms want or need to signal credibility. Therefore, it is more pronounced for private sector firms, and more pronounced when audit fee disclosure is mandatory (as has been the case in the US since 2002 and in many other countries since even earlier) which enables investors to check if auditors' fees satisfy the no-collusion condition. The importance of mandatory audit fee disclosure is indirectly supported by (DeFond, Wong and Li 2000) who find – in a study of Chinese firms at a time when audit fee disclosure was not mandatory in China[13)] – that reforms which raised accounting standards were followed by a "flight from audit quality." Firms which sought to collude with their auditors responded by switching to smaller and lower-quality auditors. This, too, is consistent with our model as smaller auditors in our setup are more willing to collude since their chances of being detected in collusion are relatively small. Along similar lines, (Wang, Wong and Xia 2008) find that state-owned enterprises in China have a higher tendency to hire small local auditors, and find

13) It was made mandatory in 2001.

evidence that one of the reasons behind this is the greater ease of colluding with these auditors. Privately owned enterprises – possibly because of credibility pressures as in our model – were on the other hand significantly less likely to use such auditors. In contrast to Wang et al., who emphasize that smaller auditors may have made more attractive collusion partners being easier to control, our model provides an additional reason why small auditors may themselves be relatively eager (compared to large ones) to engage in collusion. (Titman and Trueman 1986) contains a theoretical model in which firm insiders intending to cheat investors (on the strength of certain private information) hire a poor-quality auditor so that collusion is easier. Though our model does not deal with differences in audit quality, the number of an auditor's clients can be a rough proxy for quality, given findings of a positive association between audit size and audit quality (eg Choi et al. 2010a; Colbert and Murray 1999).

Our model implies that audit fees should be neither too small nor too large to maintain credibility (Choi, Kim and Zang 2010b). shows that audit quality tends to suffer if an audit firm is paid "abnormally high" audit fees. They argue that the abnormal audit fees are essentially bribes that induce collusive behavior on the part of auditors.

Multiple Collusions or Multiple Clients?

In the proof given above, we show that our results hold even when the rewards from collusion are maximized by the auditor colluding simultaneously with all his clients and receiving bribes from all of them (we also of course show – in an appendix – that such simultaneous collusion will be, under certain assumptions, the best course for the auditor to pursue if he is to collude). However, simultaneous collusions are by no means necessary for our results. In the appendix, we prove that they continue to hold even if the auditor colludes with just a single client so long as he has other clients as well.

The single collusion analysis calls to mind the effectiveness of *multilateral* punishments for infractions in two-player interactions as in (Greif 1991). Even without conscious multilateral punishment by the client firms, the auditor's reputation, once lost, ensures that none of them find it worthwhile to hire him in future.

This emphasis on the reputation of an independent external

auditor and on the related role played by a multiplicity of clients on collusion incentives distinguishes our work from that of previous authors: for example (Kofman and Lawarree 1993), and (Khalil and Lawarree 2006), essentially model side payments to an *internal* auditor while the external auditor is assumed to be honest. Hence, the issue of the auditor's reputation or the number of clients does not play a role[14] (Baiman, Evans and Nagarajan 1991). do not model the incentives for collusion, unlike us, but allow nature to determine whether self-enforcing collusive arrangements can prevail.[15]

Why does Auditing Relax the Credibility Ceiling?

We have shown that auditing relaxes the credibility ceiling, despite the possibility of collusion between an auditor and his clients. We have already shown that each auditor will have multiple clients. In the case of collective or multiple collusion, it is crucial that the probability of his being caught colluding increases with the number of clients he colludes with, or, equivalently, that the public signal is imperfectly correlated across firms. Because of this, an auditor has to worry about facing the penalty for collusive behavior for a greater range of parameters then does a single firm attempting to cheat in an auditor's absence. Therefore, the auditor enforces honest behavior for a greater parameter range. The imperfect correlation of the public signal across firms becomes important given the fact that the auditor in general prefers to collude with all his clients simultaneously. We must emphasize again that our results are not dependent on this simultaneous collusion, as proved in the appendix.

Multiplicity of transactions ("diversification") and imperfect correlation of the rewards and penalties from them have of course been at the heart of many phenomena in corporate governance. Managers working on multiple independent projects can be punished for neglecting a project by withholding from them the returns of their other projects without running into the constraint

14) There are many other differences from our framework including the facts that the principal pays the auditors in their models, and that the effect of auditing on entry into entrepreneurship is not modeled, as the set of firms is exogenous.

15) Another related paper is (Tirole 1986), which, though it does not deal explicitly with auditing, considers side payments in a principal-supervisor-agent hierarchical relationship. His focus is on the optimal length of such relationships.

of limited liability: this induces them to perform better (Laux 2001). Debt-financed bankers monitoring a diversified asset portfolio are likelier to be able to repay their debts: so they can expect a full-liability payoff on their monitoring efforts – a fact that improves their incentive to monitor (Cerasi and Daltung 2000; Diamond 1984). In our model, multiplicity of clients and imperfect correlation of the risk of detection of collusions facilitate the punishment of colluders, improve the incentives for honest auditing and increase the credibility of auditing as a commitment device.

Discussion: credible auditing, entrepreneurial entry and takeoff

Auditing relaxes the credibility constraint and reduces the wealth requirement for entry. This raises both the supply and demand curves for capital with the following consequences for the different equilibria specified.

First, the new interior equilibrium will replicate the old ratio of outsider to entrepreneurial capital s* (determined by the wealth distribution independently of all other factors) but at a higher payoff D for investors. Entry requirements (determined by s*) are invariant, so is the set of firms. Since all available capital was fully invested in the industry and remains so after the change, there is no impact on output, only a redistribution from firms to investors and auditors.

Second, if equilibrium occurred earlier on the horizontal stretch of the supply curve, the rise in both curves will absorb some of the excess supply of capital into the industry (at the same payoff R) or all of it (at a higher payoff). In the former case, higher s* will invite more outside capital without change in its payoff, reduce entry requirements and attract more firms. Outside investors will be no worse than before, while firms benefit – incumbents from a higher s, new entrants from a rise in payoff above opportunity cost R. The Pareto improvement is possible because excess capital earlier reduced to its outside option is now in the industry, increasing its output by more than its displaced earnings. In the latter case, there is a shift to a regular interior equilibrium. s* rises, leading to new entry and more outside investment in each firm at a higher payoff D > R. Outside investors and new firms benefit; whether incumbents benefit or not depends on whether higher s compensates for the higher payout D. In any event, industry output increases. So the gains of the gainers will more than offset the losses of the losers,

and the changed scenario is optimal according to the Hicks-Kaldor-Scitovsky compensation criteria.

Finally, where no market exists, the change could create one if it raises s*(R) above ($\underline{I}/W_{max}$) – 1, increasing credibility to the point where a firm can offer investors their opportunity cost. The Pareto-improvement is clear; so is its source – the emergence of the market. This fits in with empirical evidence by (La Porta et al. 1997) that firms tend to go public in the first place only if good measures of shareholder protection are in place. This is consistent with our implication that investor protection (like auditing), by making it possible to credibly raise share-capital, determines whether firms go public or remain family enterprises.

From our discussion we can gather the following. In case 3 – where a market is created by auditing – there are clear gains to hiring an auditor, and this happens, as we discussed before, when no one in the society has very much wealth. Here credible auditing makes it possible for industry to take off. In situation 2, too an auditor adds to net social welfare. Situation 2 – that of excess supply of capital – is likely to arise when there is a strong middle class so that much wealth is owned by those not qualified to become insiders because their wealth is below the entry threshold. Thus another situation when auditing is beneficial is when there is a strong middle class. Here, credible auditing facilitates entrepreneurial entry and increases overall output. Finally, in situation 1, firms do not gain from auditing. Investors, however, do. In this situation, we could perhaps see intermediaries hired by investors instead of firm-hired auditors.

Our results suggest that a firm-hired auditor is important either when every one in the society is strongly wealth constrained, or when there is a prominent middle class. Moreover, with mandatory audit fee disclosure – backed by heavy penalties for false disclosure – large auditors are hired and result in higher economic growth in the circumstances just identified. To make auditing credible, audit fees need to be neither too large nor too small. Moreover, in these cases, ownership is likely to become more diffuse as credible auditing increases the proportion of financing that outsiders are willing to supply.

(Choi and Wong 2007) find evidence for the "strong governance" view, finding that the demand for large/high-quality auditors is greater where legal institutions are weak. Their interpretation is that

these auditors are more effective at resolving agency problems and problems caused by asymmetric information – problems which are more severe when legal institutions providing protection to investors are weak. Our results are compatible with these findings, since we find that credible auditing is in demand and is in fact most useful in countries which are relatively poor.[16)]

Our results on the relationship between diffuse ownership and credible auditing are supported by (Simunic and Stein 1987), according to whom firms which hire a big auditor (one of the then "big eight") were likely to have a lower insider stake.

Auditors dislike any rise in public transparency for two reasons. First, lower q increases the probability that auditors can do better than the public signal and so increases demand for them. Secondly, audit fees that guarantee an honest equilibrium are lower for an accurate public signal (particularly if accuracy increases speedily with the number of clients). The logic is that if the signal is inaccurate, auditors' future fees should be high for the expected loss of such fees (should collusion be exposed) to outweigh the current period bribes the auditor could extract. Auditors therefore would dislike changes such as disclosure of stock options as costs.[17)]

In the light of our model, we now briefly discuss some suggested antidotes to corporate fraud such as the reforms partly embodied in the Sarbanes-Oxley Act. One such measure is the separation of audit from non-audit activities. The idea is that this would restrict opportunities for bribery (for example through a generous investment banking mandate to buy the auditor-cum-investment banker's collusion). Certainly, a legally enforced separation of audit from non-audit activities could increase the credibility of auditing by making covert bribery difficult and facilitating detection. In terms of our model, it would raise q, resulting in a wider range of fees satisfying the conditions for honest auditing.[18)]

Although the inaccessibility of outside credit for entrepreneurs

16) We also find that they are useful in countries with a high degree of equality, but this result does not relate directly to (Choi and Wong 2007).

17) This conclusion might be modified if we assumed that auditors can detect cheating only inaccurately. In that case, they might be helped by higher q, with the public signal complementing their efforts. But the effects mentioned above would still be present.

18) Some relevant literature on the effect of the provision of non audit services on auditing firms' tendency to qualify a report includes (Wines 1994; Barkess and Simnett 1994; Craswell, Stokes and Laughton 2002).

may significantly restrict market creation, particularly when credit market imperfection is compounded by a wealth distribution with too few wealthy individuals, auditing can partially compensate for this market imperfection. Mandatory disclosure of audit fees ensures that in spite of collusion possibilities, auditing remains credible, relaxing requirements for entrepreneurial entry as long as each auditor has multiple clients. Auditors, on the other hand, dislike a too-perfect credit market that enables firms to set up business for lower values of public transparency and thus minimizes the need for auditors.

OPTIMAL CONTRACTS

Equity contracts of course are only one of the possible ways of raising capital. Consider an alternative scenario in which the firm has a menu of contracts to choose from. The most general form of contract that our liquidity-constrained firm can offer its investors is the promise to deliver min $[D_G S, G(F + S)]$ if it is lucky and min$[D_B S, B(F + S)]$ if unlucky (where $pD_G + (1 - p)D_B = D$). This is a contract that explicitly allows for the possibility of bankruptcy. Bankruptcy is possible if $D_B S > B(F + S)$. On the other hand, if $D_B S \leq B(F + S)$ – implying $s \leq s_l = B/(D_B - B)$ – bankruptcies are impossible and any claim of bankruptcy will be legally barred.

The possibility of bankruptcy raises the specter of false bankruptcies. The entrepreneur could claim misfortune even when he has been lucky, distribute $B(F + S) < D_B S$ to his investors and decamp with the spoils $(D_G - B)S - BF$. (Bankruptcy implies closure of the firm and loss of subsequent insider profit: so false bankruptcies are unprofitable if the present value of this loss exceeds its one-time cheating gain – if

$$\frac{\delta(H - D)(F + S)}{1 - \delta} \geq (D_G - B)S - BF$$

$$\text{or } s \leq \frac{\delta(H - D) + (1 - \delta)B}{(1 - \delta)(D_G - B) - \delta(H - D)} = s_b.$$

Of course, this implies a meaningful limit on s only if $s_b > 0$, which occurs if and only if

$$(1-\delta)(D_G - B) > \delta(H - D).$$

If $s \le s_b$ or if $D_BS \le B(F + S)$, there will be no false bankruptcies. However, firms could still cheat by misrepresenting the fortunes of the business and offering investors D_BS instead of their rightful dues D_GS, risking detection by the public signal and loss of future income. This variety of cheating would be unprofitable if and only if

$$\frac{q\delta(H-D)(F+S)}{1-\delta} \ge (D_G - D_B)S$$

$$\text{or } s \le \frac{\delta q(H-D)}{(1-\delta)(D_G - D_B) - \delta q(H-D)} = s_d.$$

An optimal contract is one that maximizes s while eliminating false bankruptcies as well as the incentive to dishonestly offer investors D_BS instead of D_GS. Bankruptcy would be impossible if the contract itself provides that, in the event of bad luck, the entire proceeds of the firm should go to the outside investor: $D_BS = B(F + S)$. This is indeed the maximum guarantee against misfortune that a liquidity-constrained firm can credibly promise the outside investor: it maximizes D_B subject to credibility. If $\max D_B = B(F + S)/S \le D \le D_G$, the implication is that it minimizes the cheating premium $(D_G - D_B)$ for any given D – and thereby maximizes s_d. It follows that this is the optimal contract whenever $B(F + S)/S \le D$. Given that investors are to receive B(F + S) under bad luck, the contract must provide for a good luck payout that ensures an expected return D:

$$pD_GS + (1-p)B(F+S) = DS$$
$$\text{or } D_GS = [DS - (1-p)B(F+S)]/p.$$

With D_G and D_B thus determined, $(D_G - D_B)S$ reduces to $\{DS - B(F + S)\}/p$, yielding

$$s \le \hat{s} = \frac{\delta pq(H-D) + (1-\delta)B}{(1-\delta)(D-B) - \delta pq(H-D)}.$$

Thus, the optimal contract also imposes a ceiling on s that is a decreasing function of D.

The picture is rather different if $B(F + S)/S > D$. With a contract that offers investors $B(F + S)$ in the event of bad luck, this configuration tempts the firm to cheat the investor in the diametrically opposite way – by offering the smaller good luck payoff when in fact it has been unlucky. However, $B(F + S)/S > D$ is equivalent to $s < B/(D – B)$, s is a variable controlled by the firm and the firm's profits are an increasing function of s (if $D < H$).[19] Since, $s < B/(D – B)$ also undermines its credibility by creating moral hazard, the firm has every incentive to increase s to the level $s \geq B/(D – B)$, at which point the firm will find it worthwhile to offer the optimal contract described above.

The credibility limit for the pure equity contract is $s = s^*$. That for a pure debt contract is determined by the no-bankruptcy conditions – either $s \leq s_l$ or $s \leq s_b$, which reduce, for this contract (in which $D_B=D_G=D$), to $s \leq B/(D – B)$ and $s \leq \{\delta(H – D) + (1 – \delta)B\}/\{(1 – \delta)(D – B) – \delta(H – D)\}$ respectively. All these limits are lower than $\hat{s}$, when $D< 2B$ –when pure equity and debt contracts are dominated by the contract described in this section. However for $D > 2B$, pure debt contracts dominate as s_b is then higher than $\hat{s}$. In all cases, however, there is a limit to the size of the firm, proportional to the personal wealth of the entrepreneur. All our qualitative results will therefore continue to hold.

(Gale and Hellwig 1985) model a costly state verification problem where firms have the potential of cheating their investors. Investors (often big entities like banks) can verify the state of nature, but verification is costly. The paper recommends a debt contract (which being state-invariant, leaves the firm with less scope for cheating by exploiting its private knowledge of the state) with the proviso that declarations of bankruptcy should be followed by costly state verification by the lender. Beyond one period, however, problems of renegotiation-proofness arise. This section shows that in spite of our focus on equity contracts, in principle a very similar analysis is applicable to a debt contract as well. In that event, firms could exploit their asymmetric information about the state of nature to declare bankruptcy and give the investors a cash flow consistent with bad luck. However, the same firms could, unless caught by the public signal, again be refinanced by investors who believe that

19) $D > H$ is incompatible with the existence of the firm since the entrepreneur would then prefer to become an outside investor.

the firm genuinely experienced bad luck. If the public signal is sufficiently imprecise, investors know beforehand that the firms face too strong a moral hazard, and do not finance the industry.

EXTENSIONS

Our model suggests that the insiders' temptation to cheat is an increasing function of the ratio of outsider to insider capital, but that auditing moderates this relationship, making a higher ratio consistent with honesty. This is reflected in international differences in shareholding patterns. In the US, for example, most firms are widely held: s is higher than in countries dominated by family-owned firms or those where large blocks of capital are owned by other large companies, or banks – well represented on the board of directors and therefore "insiders." Does this difference reflect differences in the auditing framework? Is firm ownership dispersed where better auditing safeguards small outside shareholders? We find much empirical support for this view. (La Porta, Lopez-de-Silanes and Shleifer 1999) find that firms are widely held only in countries with strong shareholder protection (Shleifer and Wolfenzon 2002) comes to similar conclusions. (Simunic and Stein 1987) find that insider stake is lower in firms that hire big auditors, and (Francis and Wilson 1988) find that firms are more likely to switch to a bigger (better) auditor when ownership structure is more diffuse. In countries with weak shareholder rights, one often observes family owned firms. This fits in with family financing and underdeveloped share markets wherever shareholder protection (eg. a better auditing framework or better transparency) is inadequate. We have already mentioned (La Porta et al. 1997) which finds that firms tend to go public in the first place only if good measures of shareholder protection are in place.

Of course, a widely held shareholding pattern also implies that when auditing is flawed, there is a great risk of investors being cheated. In our model shareholders are aware of the aggregate ratio of outsider to insider capital and can make their investment decision accordingly. However if this information could be kept secret – for example if insiders secretly divest (as in Enron), raising the ratio of outsider to insider financing – outside shareholders become more vulnerable.

Appendices

Appendix A: The Zero Savings Assumption

One possible justification for the no-savings assumption in our model is as follows. With risk-neutrality and a constant rate of time-preference, the intertemporal utility function can then be written as

$$U = \sum \delta^t c_t$$

where c_t is consumption in the t-th period. The net utility increment from a one-period deferment of a unit of t-th period consumption is then

$$\delta^t[-1 + \delta(1 + r_t)]$$

where r_t is the rate of return to capital in the t-th period. With risk-neutrality, savings no longer smoothe consumption. They now reflect only the difference, if any, between the rates of time-preference and of return on capital. When these are independent of the level of consumption, savings will have an all-or-nothing character. If capital can be consumed without limit and time preference is higher than the rate of return, all wealth will be dissipated in the first period. On the other hand, if the rate of return is higher, all income will be saved and consumption deferred indefinitely. Savings will be precisely zero if (1) capital is not consumable (a standard assumption, see (Bernanke and Gertler 1989)) and (2) time preference exceeds the rate of return.[20] In our model, the highest rate of return is H: a sufficient condition for zero savings therefore is $H < (1 - \delta)/\delta$. Such a restriction is not inconsistent with any of our results.

One could of course question the origin of what wealth there is. Where did it come from if there are no savings? Here, we must resort to 'manna from heaven' assumptions. All wealth could be land, as in some banana republic where the consumption good is too perishable to be stored. Alternatively, in an industrial economy,

---

20) No one can dissave by trading capital for output, since, if one wishes to dissave, so will everyone else – so that the potential dissaver cannot find anyone to trade with.

wealth could be machinery, received by the country through foreign aid or as war reparations. Our essential purpose of course is a focus on the problem of cheating independently of the level or distribution of wealth; and all we need for this purpose is that the zero-savings assumption should be self-consistent, not that it should be realistic.

Appendix B: The Optimality of Multiple Collusions

An auditor who offers to collude with a single client can extract at most maximum the latter's maximum gain from cheating

$$X = \frac{DSL}{H} - q(1)P_h + \frac{q(1)\delta V}{1-\delta}$$

However, if exposed, he will lose the expected value of the fees paid by all of his clients – q(1)Nδ$V/(1 – δ). Denote this by q(1)Y where Y is the discounted value of audit fees from all N clients. If, on the other hand, he offered to collude with all his clients, and this offer is accepted by all, his potential income from bribes would be multiplied by N, while the risk of exposure would increase from q(1) to q(N). (We have proved in the text that an offer acceptable to one client is acceptable to all). His income expectation from multiple collusion is

$$NX - q(N)Y$$

His income from single collusion is

$$X - q(1)Y.$$

A sufficient (but not necessary) condition for him to prefer multiple collusion is

$$Nq(1) \geq q(N).$$

This condition is sufficient because if it holds, multiple collusion yields more profits for the auditor than N times the profits from single collusion (and N cannot be less than one). An interpretation of the sufficient condition is that the risk of exposure should not increase more than additively as the number of firms increases. While this seems highly plausible, multiple collusion can be optimal even under weaker conditions.

Appendix C: The Single Collusion Case

In the previous appendix we have derived conditions for an auditor to prefer multiple collusion to single collusion. However, even if the auditor preferred to collude with a single client, he would still end up relaxing the credibility ceiling, provided he has more than one client. We demonstrate this below.

In the case of collusion with a single client, the auditor's probability of detection is q(1), the same that a firm faces; however if detected he loses the future fees from all his N clients, though he is bribed by just one. This has to do with reputation effects: once the auditor is exposed as a cheat firms find it worthless to hire him as investors no longer trust him. In this case the no-collusion constraint becomes:

$$\frac{Nq(1)\delta V}{1-\delta} > \frac{DSL}{H} - q(1)P_h + \frac{q(1)\delta V}{1-\delta} \tag{19}$$

The left hand side shows the auditor's total discounted losses from all his N clients in the event of exposure by the public signal: the right hand side is the maximum bribe offered to him by the firm with which he is colluding. The bribe remains the same as in the multiple collusion case. (19) in combination with the upper limit on audit fees gives us the following range for audit fees:

$$P_h - \frac{DSL}{H} > \frac{\delta V}{1-\delta} > \frac{\frac{DSL}{H} - q(1)P_h}{(N-1)q(1)} \tag{20}$$

Following the same logic as in the multiple collusion case, firms realize that they can drive down the audit fee to its lower limit, which is this time a different one,

$$V^* = \frac{1-\delta}{\delta}\frac{DSL/H - q(1)P_h}{(N-1)q(1)},$$

and that this fee is decreasing in N: therefore each hires an auditor with as many clients as possible driving N up to its technological maximum of L. Meanwhile for the range in (20) to be non-empty we require

$$Nq(1)P_h > [1 + (N-1)q(1)]\frac{DSL}{H} \tag{21}$$

We can easily check that as long as N = L > 1, this range permits honesty in a parameter range where it would not have been possible in the absence of an auditor. We may also check that the new credibility ceiling on s becomes

$$s^{m1} = \frac{Nq(1)\delta H(H-D)}{DL(1-\delta)[1+(N-1)q(1)] - Nq(1)\delta H(H-D)} \tag{22}$$

We can show that the condition that this be greater than s*, the credibility ceiling without an auditor, boils down to:

$$s^{m1} > s^{\cdot}$$

iff

$$(1-q(1))(N-1)q(1)(1-\delta)DL > 0$$

which always holds for a less then perfectly accurate public signal, provided N = L > 1.

REFERENCES

Aghion, P. and J. Williamson (1998), *Growth, Inequality and Globalization*, Cambridge University Press.

Baiman, S., J. Evans and N. Nagarajan (1991), "Collusion in Auditing," *Journal of Accounting Research*, 25, 217-244.

Banerjee, A. V. and A. F. Newman (1993), "Occupational choice and the process of development," *Journal of Political Economy*, 101, 274-298.

Banerjee, A. V and E. Duflo (2003), "Inequality and Growth : What can the Data Say?" *Journal of Economic Growth*, 8, 267-299.

Barkess, L and R. Simnett (1994), "The pricing of other services by auditors: independence and pricing issues," *Accounting and Business Research*, 24, 91-108.

Barro, R. (2000), "Inequality and growth in a panel of countries," *Journal of Economic Growth*, 5 (1), 5-32.

Benabou, R. (2000), "Unequal Societies : Income Distribution and the Social Contract," *American Economic Review*, 90, 96-129.

Bernanke, B and M. Gertler (1989), "Agency costs, net worth and Business Fluctuations," *American Economic Review*, 79, 14-31.

Bertrand, M. and A. Schoar (2006), "The role of family in family firms," *Journal of Economic Perspectives*, 20, 73-96.

Burkart, M., F. Panunzi and A. Shleifer (2003), "Family firms," *Journal of Finance*, 58, 2167-202.

Cerasi, V. and S. Daltung (2000), "The optimal size of a bank : Costs and Benefits of Diversification," *European Economic Review*, 44, 1701-1726.

Choi, J.H and T.J Wong (2007), "Auditors' Governance Functions and Legal Environments: an International Investigation," *Contemporary Accounting Research*, 24, 13-46.

Choi, J.H, F. Kim, J.B Kim and Y. Zang (2010a), "Audit office size, audit quality and audit pricing," *Auditing: a Journal of Practice and Theory*, 29, 73-97.

Choi, J.H, J.B Kim and Y. Zang (2010b), "Do Abnormally High Audit Fees Impair Audit Quality?" *Auditing: a Journal of Practice and Theory* 29, 115-140.

Colbert, G. and D. Murray (1999), "State Accountancy Regulations, Audit Firm Size and Auditor Quality: An Empirical Investigation," *Journal of Regulatory Economics*, 16, 267-285.

Craswell, A., D. Stokes and J. Laughton (2002), "Auditor Independence and Fee Dependence," *Journal of Accounting and Economics*, 33, 253-75.

Cunningham, L. (2006), "Too Big to Fail : Moral Hazard in Auditing and the Need to Restructure the Industry before it Unravels," *Columbia Law Review*, 106, 1698-1748.

DeFond, M., T.J Wong and S. Li (2000), "The impact of improved auditor independence on auditor market concentration in China," *Journal of Accounting and Economics*, 28, 269-305.

Deininger, K. and L.Squire (1998), "New ways of looking at old issues: inequality and growth," *Journal of Development Economics*, 57(2), 259-87.

Diamond, D. (1984), "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, 51, 393-413.

Faccio, M. (2006), "Politically Connected Firms," *American Economic Review*, 96, 369-386.

Francis, J.R and E.R Wilson (1988), "Auditor Changes: a Joint Test of Theories Relating to Agency Costs and Auditor Differentiation," *Accounting Review* 63, 663-682.

Forbes, K. (2000), "A reassessment of the relationship between inequality and growth," *American Economic Review*, 90 (4), 869-87.

Gale, D. and M. Hellwig (1985), "Incentive-Compatible Debt Contracts : the One-Period Problem," *Review of Economic Studies* , 52, 647-663.

Galor, O and J. Zeira (1993), "Income Distribution and Macroeconomics,"

Review of Economic Studies, 60, 35-52.
Green, E and R Porter (1984), "Noncooperative collusion under imperfect price information," *Econometrica*, 52, 87-100.
Greif, A. (1991), "Contract Enforceability and Economic Institutions in Early Trade: the Maghribi Traders' Coalition," *American Economic Review*, 83, 525-48.
_____ (1996), "Contracting, Enforcement and Efficiency: Economics Beyond the Law,"*Annual World Bank Conference on Development Economics*, 239-265.
Joh, S.W. (2003), "Corporate Governance and Firm Profitability: Evidence from Korea before the Crisis," *Journal of Financial Economics*, 68, 287-322.
Khalil, F. and J. Lawarree (2006), "Incentives for Corruptible Auditors in the Absence of Commitment," *Journal of Industrial Economics*, 54, 269-291.
Kofman, F. and J. Lawarree (1993), "Collusion in Hierarchical Agency," *Econometrica*, 61, 629-656.
La Porta, R, F. Lopez-de-Silanes, A. Shleifer and R. Vishny (1997), "Legal Determinants of External Finance," *Journal of Finance*, 52, 1131-1150.
La Porta, R, F. Lopez-de-Silanes, and A. Shleifer (1999), "Corporate Ownership Around the World," *Journal of Finance*, 54, 471-517.
Lal, D and H. Myint (1996), *The Political Economy of Poverty, Equity and Growth,* Clarendon Press, Oxford.
Laux, C (2001), "Limited-liability and incentive contracting with Multiple Projects," *RAND Journal of Economics*, 32, 514-526.
Lemmon, M and K. Lins (2003), "Ownership Structure, corporate governance and firm value: evidence from the East Asian financial crisis," *Journal of Finance*, 58, 1445-1468
Li, H. and H.F. Zou (1998), "Increased inequality is not harmful for growth," *Review of Development Economics* 2 (3), 318-34.
Murphy, K, A. Shleifer and R. Vishny (1989), "Income Distribution, Market Size and Industrialization," *Quarterly Journal of Economics*, 104, 537-564.
Persson, T. and G. Tabellini (1994), "Is Inequality Harmful for Growth? Theory and Evidence," *American Economic Review*, 84, 600-621.
Rajan, R. and L. Zingales (1998), "Financial dependence and growth," *American Economic Review*, 88, 559-586.
Shapiro, C. and J. Stiglitz (1984), "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, 74, 433-444.
Shleifer, A. and D. Wolfenzon (2002), "Investor Protection and Equity Markets," *Journal of Financial Economics*, 66, 3-27.
Simunic, D.A and M.T Stein (1987), *Product Differentiation in Auditing: A Study of Auditor Effects in the Market for New Issues.* The Canadian Certified General Accountants' Research Foundation.

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

Titman, S and B. Trueman (1986), "Information quality and the valuation of new issues," *Journal of Accounting and Economics*, 8, 159-172.

Wang, Q., T. J Wong and L. Xia (2008), "State ownership, the institutional environment, and auditor choice: evidence from China," *Journal of Accounting and Economics*, 46, 112-134.

Wines, G (1994), "Auditor independence, audit qualifications and the provision of non-audit services: A Note," *Accounting and Finance*, 34, 75-86.

Received March 31, 2012
Revision Received June 20, 2012
Accepted August 7, 2012