Review of Economic Studies (1984) LI, 393–414 0034-6527/84/00280393$02.00

Financial Intermediation and Delegated Monitoring

DOUGLAS W. DIAMOND
University of Chicago

This paper develops a theory of financial intermediation based on minimizing the cost of monitoring information which is useful for resolving incentive problems between borrowers and lenders. It presents a characterization of the costs of providing incentives for delegated monitoring by a financial intermediary. Diversification within an intermediary serves to reduce these costs, even in a risk neutral economy. The paper presents some more general analysis of the effect of diversification on resolving incentive problems. In the environment assumed in the model, debt contracts with costly bankruptcy are shown to be optimal. The analysis has implications for the portfolio structure and capital structure of intermediaries.

INTRODUCTION

This paper develops a theory of financial intermediation based on minimum cost production of information useful for resolving incentive problems. An intermediary (such as a bank) is delegated the task of costly monitoring of loan contracts written with firms who borrow from it. It has a gross cost advantage in collecting this information because the alternative is either duplication of effort if each lender monitors directly, or a free-rider problem, in which case no lender monitors. Financial intermediation theories are generally based on some cost advantage for the intermediary. Schumpeter assigned such a "delegated monitoring" role to banks,

> ... the banker must not only know what the transaction is which he is asked to finance and how it is likely to turn out but he must also know the customer, his business and even his private habits, and get, by frequently "talking things over with him", a clear picture of the situation (Schumpeter (1939), p. 116).

The information production task delegated to the intermediary gives rise to incentive problems for the intermediary; we can term these delegation costs. These are not generally analysed in existing intermediation theories, and in some cases one finds that the costs are so high that there is no net advantage in using an intermediary. Schumpeter made a similar point, although he did not consider incentives explicitly:

> ... traditions and standards may be absent to such a degree that practically anyone can drift into the banking business, find customers, and deal with them according to his own ideas.... This in itself... is sufficient to turn the history of capitalist evolution into a history of catastrophes (Schumpeter (1939), p. 116).

This paper analyses the determinants of delegation costs, and develops a model in which a financial intermediary has a *net* cost advantage relative to direct lending and borrowing.

Diversification within the intermediary is key to the possible net advantage of intermediation. This is because there is a strong similarity between the incentive problem between an individual borrower and lender and that between an intermediary and its

depositors. The possibility of diversification within the intermediary can make the incentive problems sufficiently different to make it feasible to hire an agent (the intermediary) to monitor an agent (the borrower). Diversification proves to be important even when everyone in the economy is risk neutral.

This model is related to two literatures. It relates to the single agent-single principal literature (e.g. Harris–Raviv (1979), Holmström (1979) and Shavell (1979)) which develops conditions when monitoring additional information about an agent will help resolve moral hazard problems. The analysis here extends this to costly monitoring in a many principal setting, where principals are security holders of a firm or depositors in an intermediary. The other related literature is that of financial intermediation based on imperfect information. Several interesting papers analyse the gross benefits of delegating some informational task to an intermediary without presenting explicit analysis of the costs and feasibility of this delegation (e.g. Leland–Pyle (1977) and Chan (1982)). In addition to developing a model in which overall feasibility of financial intermediation is analysed, we briefly apply our results to determine conditions when intermediation is feasible in the Leland–Pyle model.

The basic model developed is of an ex-post information asymmetry between potential lenders and a risk neutral entrepreneur who needs to raise capital for a risky project. In this environment, debt is shown to be the optimal contract between an entrepreneur and lenders. Because of the wealth constraint that an entrepreneur cannot have negative consumption (pay lenders more than he has), the debt contracts with which the entrepreneur can raise funds involve some costs. As an alternative to incurring these costs, it is possible for lenders (who contract directly with the entrepreneur) to spend resources monitoring the data which the entrepreneur observes. In the class of contracts written directly between entrepreneurs and lenders, the less costly of these two is optimal. However, the cost of monitoring may be very high if there are many lenders. If there are m outside security holders in a firm and it costs $K > 0$ to monitor, the total cost of direct monitoring is $m \cdot K$. This will imply either a very large expenditure on monitoring, or a free rider problem where no securityholder monitors because his share of the benefit is small. The obvious thing to do is for some securityholders to monitor on behalf of others, and we are then faced with analysing the provision of incentives for delegated monitoring.

There are many methods by which delegated monitoring might be implemented. We assume that the information monitored by a given person cannot be directly observed without cost by others. The analysis here focuses on a financial intermediary who raises funds from many lenders (depositors), promises them a given pattern of returns, lends to entrepreneurs, and spends resources monitoring and enforcing loan contracts with entrepreneurs which are less costly than those available without monitoring. The financial intermediary monitors entrepreneurs' information, and receives payments from the entrepreneurs which are not observed by depositors.

An example of useful costly information in a loan contract is a covenant which is costly to monitor. A common covenant is a promise that the firm's working captial will not fall below some minimum, unless "necessary for expansion of inventory". (See Smith–Warner (1979).) If it is costly to determine whether a shortfall is "necessary", and each of the bondholders has to incur this cost to enforce the contract, the contract using costly information is unlikely to be used if the number of bondholders is large. A contract specifying an uncontingent working capital requirement might be substituted, when the contingency would have been specified if there had been a single principal. In practice, loan covenants in bank loan contracts specify coarse contingencies which define

a "default". Conditional on such a default, the intermediary monitors the situation and uses the information to re-negotiate the contract with new interest rates and contingent promises. A financial intermediary must choose an incentive contract such that it has incentives to monitor the information, make proper use of it, and make sufficient payments to depositors to attract deposits. Providing these incentives is costly, but we show that diversification serves to reduce these costs. As the number of loans to entrepreneurs with projects whose returns are independent (or independent conditional on observables) grows without bound, we show that costs of delegation approach zero, and that for some finite number of loans financial intermediation becomes viable, considering all costs.

Financial intermediaries in the world monitor much information about their borrowers in enforcing loan covenants, but typically do not directly announce the information or serve such an auditor's function. The intermediary in this model similarly does not announce the information monitored from each borrower, it simply makes payments to depositors. We show that debt is the optimal contract between the intermediary and depositors. The result that the delegation costs go to zero implies that asymptotically no other delegated monitoring structure will have lower costs. If there is an independent demand by entrepreneurs for monitoring without disclosure of the information monitored, for example to keep competitors from learning the information as suggested by Campbell (1979), then well diversified financial intermediaries can provide it (in addition to simple monitoring services) at almost no cost disadvantage.

Diversification is key to this theory, and it is interesting that because of the wealth constraint, diversification is important despite universal risk neutrality. To develop a more general intuition into the role of diversification, some analysis is presented of a related model with risk averse agents but no wealth constraint. Two types of diversification are considered in the context of two alternative financial intermediary models; one is the traditional diversification by sub-dividing independent risks, while the other is diversification by adding more independent risks of given scale. The latter is what Samuelson (1963) has termed a "fallacy of large numbers", because it does not always increase expected utility. This section may be of independent interest because it provides some conditions when the fallacy of large numbers is not a fallacy.

The basic model is outlined in Section 2. Delegated monitoring by a financial intermediary in the context of the basic model is analysed in Section 3. Section 4 explores the extension of the basic model to risk averse agents. Section 5 applies the analysis of section 4 to the model of Leland–Pyle. Section 6 concludes the paper.

2. A SIMPLE MODEL OF FIRM BORROWING

A model of risk neutral entrepreneurs who need to raise capital to operate a large investiment project is used to capture many of the aspects of the agency relationship between commerical borrowers and lenders. We specify a simple environment, and characterize optimal direct contracts between borrower and lender.

There are N entrepreneurs indexed by $i = 1, \ldots, N$ in the economy. For the balance of Section 2, we examine one of them, and do not use the index. The entrepreneur is endowed with the technology for an indivisible investment project with stochastic returns. The scale of inputs for the project greatly exceeds both his personal wealth and the personal wealth of any single lender. For simplicity, the entrepreneur's wealth is zero. Assume a one good economy with all consumption at the end of the period. The project requires inputs of the good today, and will produce output in one period. Normalize the required initial amount of inputs to one. The expectation of the output that will be

produced at the end of the period exceeds R, the competitive interest rate in the economy. Therefore, the project would be undertaken if the risk neutral entrepreneur had available to him enough capital inputs.

The other investors in the economy are also risk neutral: call them lenders. To undertake the project, the entrepreneur must borrow sufficient resources from them to operate it at its scale of one. Because the interest rate is R, i.e. the lenders have access to a technology which will return R per unit of input, the entrepreneur must convince potential lenders that the rate of return which he will pay to them has an expected value of at least R. Each lender has available wealth of $1/m$, thus the entrepreneur must borrow from $m > 1$ lenders. The capital market is competitive—if convinced that their expected return equals or exceeds R (R/m per lender), lenders will make the loan.

Let the total output of the project be the random variable $\tilde{y}$. Assume that $\tilde{y}$ is bounded between zero and $\bar{y} < \infty$. The entrepreneur and all lenders agree on the probability distribution of $\tilde{y}$, in particular all agree that $E_{\tilde{y}}(\tilde{y}) > R + K$ (where $K > 0$ and is defined below) and that $y = 0$ is possible. The realization of $\tilde{y}$ does not depend on any actions of the entrepreneur.

A simple information asymmetry is introduced which will make the loan contracting problem non-trivial. The realization of $\tilde{y}$ is freely observed only by the entrepreneur. With output observed by the entrepreneur alone, he must be given incentives to make payments to lenders. At the end of the period, he will pay a liquidating dividend. It is always feasible for him to claim a very low value of y, and keep for himself the difference between the actual value and what he pays the others.

Let $z \geqq 0$ be the aggregate payment which the entrepreneur pays to the m lenders. If the realization of output is $\tilde{y} = y$, he then keeps $y - z$ for himself. Because consumption cannot be negative the payment which he pays cannot feasibly exceed y (plus any personal wealth he might have, assumed here to be zero). To induce the entrepreneur to select a value of $z > 0$, he must be provided with incentives. To raise captial to undertake the project, lenders must believe that the expectation of the value of z which he will select is at least R. The entrepreneur must choose an incentive contract which depends only on observable variables and makes lenders anticipate a competitive expected dividend. The only costlessly observable variable is the payment z itself.

Lenders know the distribution of $\tilde{y}$, and know that the entrepreneur chooses the payment z which is best for him given a realization $\tilde{y} = y$, and that $z \in [0, y]$. If y exceeded R with probability one, then a full information optimal contract would be feasible—the risk neutral entrepreneur would offer an uncontingent payment of R. (See Harris–Raviv (1979).)

It might appear that the assumption that $y = 0$ is a possible outcome of the project rules out any borrowing, because $z = 0$ must be feasible, and it does not appear incentive compatible for an entrepreneur to choose a payment $z > 0$ when he can choose $z = 0$ and retain the rest. However, we will allow contracts with non-pecuniary penalties: penalties where the entrepreneur's loss is not enjoyed by the lenders. This allows the agent's utility function to be defined over negative values of its domain without allowing negative consumption to "produce" goods. We will see that these penalties are best interpreted as bankruptcy penalties. Some examples include a manager's time spent in bankruptcy proceedings, costly "explaining" of poor results, search costs of a fired manager, and (loosely) the manager's loss of "reputation" in bankruptcy. Physical punishment is a less realistic example. Projects which could not be undertaken at all without the penalties can be operated using the penalties.

The optimal contract maximizes the risk neutral entrepreneur's expected return, given a minimum expected return to lenders of R. Let the function ϕ, from the non-negative reals to the non-negative reals, be the non-pecuniary penalty function, which depends on z, the payment to lenders selected by the entrepreneur. Assume that if the entrepreneur is indifferent between several values of z, he chooses the one preferred by the lender. The optimal contract with penalties $\phi^*(\cdot) \geqq 0$ solves[1]

$$\max_{\phi(\cdot)} E_{\tilde{y}}[\max_{z \in [0,\tilde{y}]} \tilde{y} - z - \phi(z)] \tag{1a}$$

$$\text{Subject to} \quad z \in \arg\max_{z \in [0,y]} y - z - \phi(z) \tag{1b}$$

and

$$E_{\tilde{y}}[\arg\max_{z \in [0,\tilde{y}]} \tilde{y} - z - \phi(z)] \geqq R, \tag{1c}$$

where the notation "arg max" denotes the set of arguments that maximize the objective function that follows.

Proposition 1. *The optimal contract which solves* (1) *is given by* $\phi^*(z) = \max(h - z, 0)$, *where* h *is the smallest solution to*

$$(P(\tilde{y} < h) \cdot E_{\tilde{y}}[\tilde{y} | y < h]) + (P(\tilde{y} \geqq h) \cdot h) = R. \tag{2}$$

That is, it is a debt contract with face value h *and a non-pecuniary bankruptcy penalty equal to the shortfall from face,* h, *where* h *is the smallest face value which provides lenders with an expected return of* R.

Proof. Given $\phi^*(z)$,

$$\arg\max_{z \in [0,y]} y - z - \phi^*(z) = \begin{cases} y & \text{if } y < h \\ h & \text{if } y \geqq h. \end{cases}$$

Using (2), this satisfies with equality the constraint (1c) of providing a competitive return to lenders. By construction, h is the smallest number such that if the constraints $z \leqq y$ and $z \leqq h$ are satisfied, the expectation of $\tilde{z}$ is at least R. Hence, to satisfy (1c), there must exist some payment $h^+ \geqq h$ which is incentive compatible. If $z = h^+$ is incentive compatible (fulfills (1b) given contract $\phi(z)$), it must be true that

$$y - h^+ - \phi(h^+) \geqq \max_{z' \in [0,h^+]} y - z' - \phi(z')$$

or for all $z' \in [0, h^+]$,

$$\begin{aligned} \phi(z') &\geqq h^+ + \phi(h^+) - z' \\ &\geqq h + \phi(h^+) - z' \\ &\geqq h - z' \\ &= \phi^*(z'). \end{aligned}$$

The final inequality follows from the requirement $\phi(z) \geqq 0$ for all z. Combined with the result that $\phi^*(z) = 0$ for all $z \geqq h$, this implies that $\phi^*(z)$ gives the smallest penalties such that it is incentive compatible to fulfill (1c), implying that $\phi^*(z)$ maximizes (1). ‖

The necessity of a positive probability of incurring the non-pecuniary penalty means that even the optimal contract is costly. Entrepreneurs could be made better off without making lenders worse off if $\tilde{y}$ were observable. In a one entrepreneur-one lender setting,

where $\tilde{y}$ could be observed at some cost, it would be observed so long as the cost were less than the expected non-pecuniary bankruptcy penalty, $E_{\tilde{z}}[\phi^*(\tilde{z})] = E_{\tilde{y}}[\phi^*(\tilde{y})]$.

In a setting where it is not possible to make $\tilde{y}$ observable to lenders at some cost, the contracting problem is not influenced by the number of lenders—a contract with one lender who loans one unit is equivalent to a contract with m lenders who each loan $1/m$ units and the entrepreneur incurs a penalty of $\phi^*(z_j)/m$ on the basis of payments z_j to lender j. However, this is not true when costly monitoring is possible. By spending $K > 0$ in resources, a lender can observe $\tilde{y}$, but other lenders do not automatically observe $\tilde{y}$ as a result and other lenders cannot observe the payments by an entrepreneur to the lender who monitors. As a result, an entrepreneur and a given lender consider only the effect on a given lender's loan of $1/m$ units when deciding between a contract with costly penalties and one that avoids the bankruptcy penalties by costly monitoring.

We analyse the impact of an information technology which allows costly monitoring of the exact realization of output, y, for each entrepreneur. This information costs $K > 0$ for each principal to monitor, and the cost must be incurred before the output realization is known to anyone, including the entrepreneur. See Townsend (1979) for some interesting analysis of the optimal contingent monitoring policy when the decision to monitor can be made *after* the entrepreneur has made a payment to a lender. This additional complication is not introduced because given some specified probability of monitoring it would not influence our results.

If it is possible for lenders to observe the outcome at some cost, there are three types of contracting situations possible. The contract can be as described above, with no monitoring. A second possiblility is for each of the m lenders to spend resources to monitor the outcome. Thirdly, the lenders can delegate the monitoring to one or more monitoring agents. The least costly of these will be selected.

If there were a single lender so $m = 1$ (rather than $m > 1$ as we assume), monitoring would be valuable if its cost were less than the expected deadweight penalty without monitoring or $K \leqq E_{\tilde{y}}[\phi^*(\tilde{y})]$. With many lenders and direct contracting between the entrepreneur and lenders, if each lender monitors, monitoring is valuable if and only if $m \cdot K \geqq E_{\tilde{y}}[\phi^*(\tilde{y})]$. When m is large this is unlikely because each lender's loan is small. Even if this condition for valuable monitoring is satisfied, it implies a large expenditure on monitoring and some sort of delegated monitoring might be desirable in this case.

To obtain the benefits of monitoring, when m is large the task must be delegated rather than left to each individual lender. The entity doing the monitoring ("the monitor") must be provided with incentives to monitor and enforce the contract. We assume that the actions taken and the information observed by the monitor are not directly observed by the lenders. It will generally be costly to provide incentives to the monitor, and below we analyse these costs. The total cost of delegated monitoring is the physical cost of monitoring by the monitor, K, plus the expected cost of providing incentives to the monitor, which we call the cost of delegation and denote the cost per project by D. Delegated monitoring pays when

$$K + D \leqq \min [E_{\tilde{y}}[\phi^*(\tilde{y})], (m \cdot K)].$$

The costs of delegation are analysed when the monitor is a financial intermediary who receives payments from entrepreneurs and makes payments to principals.

3. DELEGATED MONITORING BY A FINANCIAL INTERMEDIARY

A financial intermediary obtains funds from lenders and lends them to entrepreneurs. Economists have tried to explain this intermediary role by arguing that the financial

intermediary has a cost advantage in certain tasks. When such tasks involve unobserved actons by the intermediary or the observation of private information, then an agency/incentive problem for the intermediary may exist. Any theory which tries to explain the role of intermediaries by an information cost advantage must net out the costs of providing incentives to the intermediary from any cost savings in producing information. Existing intermediary theories do not make this final step. We now introduce a financial intermediary between entrepreneurs and lenders (whom we call depositors from now on), and examine conditions when this intermediary function is viable considering all costs.

A financial intermediary is a risk neutral agent, with personal wealth equal to zero. The intermediary receives funds from depositors to lend to entrepreneurs and is delegated the task of monitoring the outcomes of entrepreneurs' projects on behalf of depositors. Monitoring the i-th entrepreneur costs the intermediary K units of goods.[2] Depositors can observe the payment they receive from the intermediary, but cannot observe the project outcomes, payments by entrepreneurs to the intermediary, or the resources expended by the intermediary in monitoring the outcomes.

Each entrepreneur's project requires one unit of initial capital. Each depositor has available capital of $1/m$, as in Section 2. An intermediary which contracts with N entrepreneurs has $m \cdot N$ depositors.

To analyse the conditions when intermediation is beneficial (when the monitoring cost savings exceed the delegation costs of providing incentives) we must first characterize the delegation costs. If the intermediary could monitor at no cost, it could enforce contracts with entrepreneurs which imposed no deadweight bankruptcy costs on them. However, there would remain an incentive problem for the intermediary, because the payments it receives from entrepreneurs are not observed by depositors. The intermediary could claim that payments from entrepreneurs were low, and pay a small amount to depositors. We now extend the results in Section 2 to analyse the optimal contract to provide incentives for an intermediary to make payments to depositors. We later show that it provides incentives to monitor as well.

Let us re-introduce the subscript i on the outcome y_i of the i-th entrepreneur. For $i = 1, \ldots, N$, the $\tilde{y}_i$ are distributed independently and all are bounded below by zero and above by the real number $\bar{y}$. The probability distribution functions of the $\tilde{y}_i$ are common knowledge to all. Let $g_i(\cdot)$ be the non-negative real valued function which is the payment to the intermediary by the i-th entrepreneur as a function of the outcome y_i, assuming the intermediary monitors y_i. Because y_i is then observed by the intermediary, this implies no deadweight penalties will be imposed on the i-th entrepreneur. If the intermediary does not monitor, it must use a contract with deadweight bankruptcy penalties, as in Section 2, but in that case there would be no reason to have an intermediary. Due to the constraint that an entrepreneur can pay only what he has, we require $g_i(y_i) \leqq y_i$. The intermediary monitoring N entrepreneurs receives total payments G_N when $\tilde{y}_1 = y_1$, $\tilde{y}_2 = y_2, \ldots, \tilde{y}_N = y_N$ equal to

$$G_N = \sum_{i=1}^{N} g_i(y_i).$$

Let $\tilde{G}_N$ be the random variable with realization G_N. It is bounded above by $\bar{G}_N$, and below by zero.

The intermediary must make total payments to depositors with expectation R per project, or $N \cdot R$ in total. Let Z_N be the total payment to depositors by entrepreneurs. The intermediary can pay only what it has, thus $Z_N \leqq G_N$. By an argument identical to that of Section 2, we see that deadweight bankruptcy penalties must be imposed on the intermediary unless the intermediary will always receive aggregate payments of at least $N \cdot R$, or $P(\tilde{G}_N \geqq N \cdot R) = 1$. Because of the constraint that entrepreneurs can pay the

intermediary at most y_i, we know $P(\tilde{G}_N \geqq N \cdot R) \leqq P(\sum_{i=1}^{N} \tilde{y}_i \geqq N \cdot R)$. Any entrepreneur, i, with $P(\tilde{y}_i \geqq R) = 1$ could finance directly, with no bankruptcy penalties, thus entrepreneurs who choose to use intermediaries will lead the intermediary to incur expected deadweight bankruptcy penalties.

Let $\Phi(Z_N)$ be the deadweight non-pecuniary penalty imposed on the intermediary when payment Z_N is made to depositors. From Proposition 1, the optimal $\Phi(Z_N)$ which gives incentives to make payments with expectation $N \cdot R$, is given by

$$\Phi(Z_N) = \max [H_N - Z_N, 0],$$

where the constant H_N is the smallest solution to

$$\{(P(\tilde{G}_N < H_N) \cdot E_{\tilde{G}_N}[\tilde{G}_N \mid G_N < H_N]) + ([1 - P(\tilde{G}_N < H_N)] \cdot H_N)\} \geqq N \cdot R.$$

With this contract in place, the expected return of the intermediary is $E_{\tilde{G}_N}[\tilde{G}_N] - H_N$; therefore the intermediary chooses monitoring expenditure to maximize $E_{\tilde{G}_N}[\tilde{G}_N]$. The intermediary uses the same decision rule in the decision to monitor as it would if its expenditure on monitoring were freely observable. This implies that it contracts only with entrepreneurs for whom the value of monitoring exceeds its physical and delegation costs, and chooses to monitor them. The minimum cost contract which provides incentives for payment to depositors also provides incentives for monitoring. This implies that the optimal contract between the intermediary and depositors is also a debt contract.

Diversification and the viability of intermediation

For a financial intermediary to be viable, three conditions must be fulfilled. The depositors must receive an expected return of R per unit deposited. The intermediary must receive an expected return net of monitoring costs and any deadweight penalties incurred which is at least zero. Finally, each entrepreneur must retain an expected return at least as high as he would by contracting directly with depositors.

Everyone in the economy is risk netural, implying that a complete description of the optimality of any feasible set of contracts is the sum of monitoring costs and expectation of total deadweight bankruptcy penalties.

A financial intermediary which contracts with one entrepreneur (and m depositors) is not viable. This follows immediately from the constraint $g_1(y_1) \leqq y_1$, that the entrepreneur can pay no more than the outcome y_1, and the constraint $Z \leqq G_1 = g_1(y_1)$, that the intermediary can pay no more than it receives from the entrepreneur. An entrepreneur incurs a deadweight penalty whenever $y_1 < h$; an intermediary with one entrepreneur who pays g_i must also incur a penalty of at least the same magnitude when $g_i \leqq y_i \leqq H_1$ (and it is necessary that $H_1 \geqq h$ to provide depositors with a competitive return). The intermediary is not viable because it incurs at least as high a deadweight cost and in addition spends resources on monitoring.

The case of one entrepreneur demonstrates the potential hazard of neglecting the costs of delegation when considering financial intermediation. The per-entrepreneur cost of providing incentives to the intermediary is reduced as it contracts with more entrepreneurs with independently distributed projects. With independent and identically distributed projects, the per-entrepreneur cost, D_N, is a monitonically decreasing function of the number of entrepreneurs, N, because deadweight penalties are incurred when returns are in the extreme lower tail, and the probability of the average return across projects being in that tail is monitonically decreasing.[3]

The argument in footnote 3 shows that for all projects with less than perfect correlation, the delegation cost for N projects monitored by a single intermediary is less than the sum of the delegation costs for monitoring proper subsets of them by several intermediaries. Increasing returns to scale from delegation cost savings is a very general result. The assumption of independence allows a stronger result. The expected delegation cost per entrepreneur monitored by the intermediary gets arbitrarily small as N, the number of entrepreneurs with independently distributed projects, grows without bound. This implies that the total cost (per entrepreneur) of providing monitoring converges to K, the physical cost of monitoring. This is Proposition 2.

Proposition 2. *The cost of delegation, per entrepreneur monitored, D_N, approaches zero as $N \to \infty$ if entrepreneurs' projects have bounded returns, distributed independently.*

Proof. Choose payment schedules $g_i = g_i(y_i)$ to the intermediary for entrepreneurs $i = 1, \ldots, N$, such that

$$E_{\tilde{g}_i}[\tilde{g}_i] = R + K + D_N, \quad \text{where } D_N > 0 \text{ is a real number.}$$

This provides the i-th entrepreneur with an expected return given by

$$E_{\tilde{g}_i}[\tilde{g}_i] - R - D_N.$$

Choose the non-pecuniary bankruptcy penalties of the intermediary as

$$\Phi_N(Z_N) = \max[(Z_N - H_N), 0]$$

where $H_N = N \cdot (R + D_N/2)$.

Given this contract, the intermediary will choose payments Z_N to depositors equal to

$$Z_N = \begin{cases} G_N & \text{if } G_N \leqq H_N \\ H_N & \text{if } G_N > H_N \end{cases}.$$

The expected return of the intermediary net of expenditure NK on monitoring is

$$E_{\tilde{G}_N}(\tilde{G}_N) - H_N - NK = [N \cdot (R + K + D_N)] - \left[N \cdot \left(R + \frac{D_N}{2}\right)\right] - (N \cdot K)$$

$$= \frac{N}{2} D_N > 0.$$

(satisfying the constraint that this be non-negative.)

The aggregate expected return to depositors is given by

$$P_N \cdot E_{\tilde{G}_N}[\tilde{G}_N | G_N \leqq H_N] + (1 - P_N) \cdot H_N \quad \text{where } P_N \equiv P(\tilde{G}_N \leqq H_N).$$

Notice that $G_N \geqq 0$ implying $E_{\tilde{G}_N}[\tilde{G}_N | G_N \leqq H_N] \geqq 0$ and that the aggregate expected rerurn of depositors is greater than or equal to:

$$(1 - P_N) \cdot H_N = (1 - P_N) \cdot N \cdot \left(R + \frac{D_N}{2}\right)$$

$$> N \cdot R \quad \text{for small } P_N > 0$$

i.e. for $P_N \in (0, (D_N/2)/(R + D_N/2))$.

There exists $N^* < \infty$ such that $P_N < \delta$ for all $\delta > 0$, by the (weak) law of large numbers, because $E_{\tilde{G}_N}[\tilde{G}_N] > H_N$. This implies that the delegation cost D_N can be made arbitrarily small for large N. ‖

Proposition 2 demonstrates the key role of diversification in the provision of delegated monitoring. The intermediary need not be monitored because it takes "full responsibility" and bears all penalties for any short-fall of payments to principals. The diversification of its portfolio makes the probability of incurring these penalties very small and allows the information collected by the intermediary to be observed only by the intermediary.

Proposition 1 characterized the optimal incentive compatible mechanism for financial intermediation, and this is the optimal incentive compatible mechanism with "privacy". It was the optimal mechanism when the agent monitoring entrepreneurs was constrained not to announce the values of the project outcomes he observed and could only use the information privately to enforce his contract with each entrepreneur. Proposition 2 shows that financial intermediation is, asymptotically, the optimal incentive compatible mechanism for financing entrepreneurs' projects, without imposing the constraint of "privacy".

If the number of entrepreneurs monitored is $N=1$, then delegation costs are so large that intermediation is never viable. If $N\to\infty$, then expected delegation costs approach zero, and intermediation is viable whenever direct monitoring pays. There exists some $N>1$ at which intermediation becomes just viable (when $D_N \leqq \min[E_{\tilde{y}}[\phi^*(\tilde{y})], m\cdot K]$). If the assumption is made that each entrepreneur's project has the same variance, then the expected delegation costs are a monotonically decreasing function of N. This leads to increasing returns to scale due to diversification, but asymptotic constant returns to scale because expected delegation costs per project are bounded below by zero, and they may be small for moderate values of N.

The incentive contract is debt with bankruptcy penalties and high leverage. Asymptotically, the debt is riskless (as $D_N\to 0$). The leverage is high, as the face value of the debt is $H(N)=N\cdot(R+D_N/2)$, while the expected future value of the intermediary (including value of the debt) is $N\cdot(R+D_N+K)$.

The importance of the diversification is not simply a way for principals to hold well-diversified portfolios. Principals are risk neutral, and are not made directly better off by the diversification. Diversification within the financial intermediary organization is important, and cannot be replaced by diversification across intermediaries by principals.

Correlated returns of entrepreneurs

The assumption of independently distributed project returns across entrepreneurs is quite strong. It can be weakened somewhat. Instead of independence, assume that entrepreneur's project returns depend on several common factors which are observable. Factors might include GNP, interest rates, input prices, etc. Since these are observable, they can be used as the basis for contingent contracts. There might exist futures markets for these variables, and the financial intermediary could hedge changes in these factors in those markets. An example is a bank's hedging of interest rate risk using interest rate futures. If there are not active futures markets, then the intermediary can write contracts with depositors which depend on the values of these factors, rather than taking responsibility for all risks. An example of this is matching the maturity of assets and liabilities by banks, which places all interest rate risk on depositors. In either case, the intermediary retains responsibility for (and potentially fails as a result of) all risks which are not observable.

The result of Proposition 2, that $D_N\to 0$ as $N\to\infty$ follows given this alternative assumption in place of independence. This is stated in the following corollary.

Corollary to Proposition 2. *If it is common knowledge that the returns of the projects of entrepreneurs $i = 1, \ldots N$, are given by*

$$\tilde{y}_i = \sum_{j=1}^{M} [\beta_{ij} \cdot \tilde{F}_j] + \tilde{\varepsilon}_i$$

where the $\tilde{F}_j$ are observable ex post, the $\tilde{\varepsilon}_i$ are independent and bounded and $E[\tilde{y}_i] > R + K$, then the result of Proposition 2 follows.

Proof. Choose $g_i(y_i) = \alpha_i \cdot y_i$ where

$$\alpha_i = \frac{R + K + D_N}{E_{\tilde{y}}[\tilde{y}_i]}.$$

Let the penalty contract be either

$$\phi(Z) = Z + [\sum_{i=1}^{N} \sum_{j=1}^{M} \alpha \cdot \beta_{ij} \cdot F_j] - H(N)$$

where

$$H(N) = \left[N \cdot \left(R + \frac{D_N}{2} \right) \right] - E[\sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \cdot \beta_{ij} \cdot \tilde{F}_j],$$

or let $\phi(Z)$ be as in Proposition 2, and let the position in the futures market be $\sum_{i=1}^{N} \alpha_i \cdot \beta_{ij}$ in futures markets $j = 1, \ldots, M$. The transformed random variables are now independent, and the result Proposition 2 follows. ‖

The intermediary monitors firm specific information, which is independent across entrepreneurs, and hedges out all systematic risks. The description of the process generating project returns is consistent with the Arbitrage Pricing Theory of Ross (1976).

The intuition behind this result is that the intermediary must bear certain risks for incentive purposes, but that risks which have no incentive component because they are common information should be shared optimally.[4] There has been a debate among various bankers and bank regulators over the desirability of allowing hedging in futures markets by banks. Our analysis suggests a reason why it is desirable.

4. RISK AVERSION AND DIVERSIFICATION

Diversification proved to be important to reduce delegation costs despite universal risk neutrality because of the wealth constraint of non-negative consumption and the asymmetry of information about project outcomes. The wealth constraint gives rise to a special type of "risk aversion". In this section, we investigate the role of diversification within the intermediary when the agents within the intermediary are risk averse in the usual sense. To focus on risk sharing issues, we drop the wealth constraint to allow any promise to be made good. A complete re-analysis of the model of Section 2 is not presented. This section does not present a realistic intermediary model, but simply a further investigation of the role of diversification in reducing the costs of delegation.

The basic set-up is as in Section 2, each entrepreneur is endowed with a project with outcome $\tilde{y}_i$ which is freely observed only the entrepreneur, which has zero as a possible realization. Absent monitoring by lenders, no incentive compatible payment schedule can depend on the realization y_i, because the entrepreneur could always claim a low value occurred. For simplicity, assume that all agents in the economy, including the

entrepreneurs, are identical and risk averse. Risk aversion implies that the payment to lenders will be a constant, rather than a random amount independent of $\tilde{y}_i$, (see Holmström (1979)). This implies that, absent monitoring, the risk averse entrepreneur bears all of the risk from fluctuations in $\tilde{y}_i$. This is inconsistent with the optimal risk sharing which would occur if $\tilde{y}_i$ were observed, and this provides a potential benefit from monitoring $\tilde{y}_i$. In this risk averse setting, we could introduce other actions, e.g. effort, which the entrepreneur could privately select to give rise to a more general motivation for monitoring. This would not change the essence of our results.

We focus again on delegated monitoring by a financial intermediary. A financial intermediary raises funds from depositors who do not monitor, lends these funds to entrepreneurs, and can offer improved risk sharing with an entrepreneur because the intermediary's monitoring reduces or eliminates the incentive problem. In Section 2, we showed that an intermediary monitoring a single entrepreneur would have an incentive problem just as severe as would an entrepreneur. Almost the same result is true here. Because depositors do not monitor the intermediary and cannot observe its information, incentive compatible payments from the intermediary to depositors cannot depend on outcomes, and will be constant. It is true, however, that a single intermediary and a single entrepreneur can now share $\tilde{y}_i$ risk, but this has little to do with intermediation. Any lender who spends resources to monitor $\tilde{y}_i$ can share risk with the entrepreneur without being called an "intermediary".

For a financial intermediary in an economy where everyone is risk averse to viably provide delegated monitoring services, it must have lower delegation costs than an entrepreneur. Equivalently, since risk sharing is the issue here, a viable financial intermediary which monitors many entrepreneurs with independently distributed projects must charge a lower Arrow-Pratt risk premium for bearing the risk of an entrepreneur's project than does the entrepreneur. This will carry over to more general settings, because if the intermediary can bear risks at a lower risk premium it will generally face a less severe trade-off between risk sharing and incentives, and can thus efficiently be delegated a monitoring task.

Two types of diversification

There are two ways in which an intermediary in an economy of risk averse agents might use diversification. They correspond to two different models of an intermediary. One model increases the number of agents working together within the intermediary organization as the intermediary monitors a larger number of entrepreneurs. The second model assumes that the intermediary consists of a *single* agent who monitors a large number of entrepreneurs with independent projects.

Beginning with the first model, assume that each identical agent ("banker") in the intermediary is risk averse, and that by spending resources to monitor, each banker within the intermediary can observe the information monitored by all other bankers within the intermediary. This implies that there are no incentive problems within the intermediary. The extreme assumption that incentive problems are absent is intended to capture the idea that there may be different mechanism for controlling incentive problems within an organization. This approach is followed in Ramakrishnan–Thakor (1983), to generalize the risk neutral analysis we present in Section 2. This model leads to the traditional "risk subdividing" type of diversification. This type of diversification works because each independent risk is shared by an increasing number of bankers. For example, each risk averse agent will obtain a higher expected utility if each of N agents invests in a fraction $1/N$ of N identical independent gambles than in any single one of the gambles.

The second type of diversification, "adding risks", occurs in the second model where a single banker bears 100% of N independent risks, with diversification occurring as N grows. This is quite different from risk subdivision, because it is not a form of risk sharing at all. The total risk imposed on the agent rises with N, while with subdivision of risks it falls with N. Samuelson (1963) termed diversification by adding risks a "fallacy of large numbers", because it is not true for all risk averse utility functions that the risk aversion toward the N-th independent gamble is a decreasing function of N. Samuelson provides no analysis of conditions when this type of diversification is beneficial, and I know of none in the literature.[5] We provide a partial characterization of conditions when the certainty equivalent of a given gamble is higher (and the risk premium lower) when another independent risky gamble is also held. That is, when is the per asset certainty equivalent higher with $N=2$ than with $N=1$?

We turn first to the relatively straightforward model of diversification by subdivision of risks. Assume that it takes one banker in the intermediary to monitor one entrepreneur, and this requires an expenditure in goods of K (or that the disutility of this monitoring task is additively separable). All bankers are identical, have increasing, concave utility of wealth functions $U(W)$. By spending K to monitor their entrepreneur, each banker can also observe information monitored by the other bankers within the intermediary, implying that there is no incentive problem within the intermediary. Depositors are not assumed to be able to observe any of the information generated within the intermediary, and are paid a fixed unconditional payment of NR. As $N \to \infty$, Ramakrishnan–Thakor (1983) shows that each banker bears an arbitrarily small risk, with perfect risk sharing within the intermediary.

The interpretation of this result is that the diversification which occurs when bankers within the intermediary can share independent risks does serve to reduce the severity of its incentive problem. This occurs because the incentive problem here imposes a constraint on optimal risk sharing, and if there is improved risk sharing within the intermediary (where incentive problems may be controlled directly, or absent as assumed here), then this is analogous to reducing the risk aversion of a single agent, which reduces the tradeoff between risk sharing and the provision of incentives.

In the second model, where the intermediary consists of a single agent, diversification by adding risks is at work. The intermediary agent monitoring N loans, receives payments from each entrepreneur and bears all of the risk because he pays an unconditional return, $N \cdot R$, to depositors. The financial intermediary can provide monitoring and risk sharing services superior to an individual lender if and only if his risk aversion toward the Nth independent risk is a decreasing function of N. Put another way, when there is no wealth constraint an intermediary monitoring a single entrepreneur ($N=1$) is equivalent to direct monitoring by a lender. Intermediation becomes potentially viable when the delegation cost (equal to the risk premium here) is reduced by the centralization of monitoring to a single intermediary. This is therefore equivalent to the conditions when adding independent risks reduces per-entrepreneur risk aversion, which are the conditions when the fallacy of large numbers is not a fallacy.

To provide a partial characterization of conditions when the per-risk risk premium declines, we initially focus on the case of two risks. That is, given two bounded and independent random variables $\tilde{g}_1$ and $\tilde{g}_2$, when is the risk premium for bearing the risk of the bounded random variable $\tilde{g}_1 + \tilde{g}_2$ less than the sum of the two risk premia for bearing either risk separately. If both random variables represent payment schedules from entrepreneurs which a risk averse intermediary would voluntarily accept, both must have expectation greater than $R+K$, because the intermediary promises $N \cdot R$ to depositors and spends $N \cdot K$ on monitoring. It will ease exposition to provisionally assume

$E_{\tilde{g}_1}[\tilde{g}_1] = R + K$. In addition, define $x_i \equiv g_i - R - K$, for $i = 1, 2$. With this notation, the net effect of contracting to monitor an entrepreneur with payment schedule $\tilde{g}_1$ is equivalent to receiving the random variable $\tilde{x}_1$. (In this notation, our temporary assumption is $E_{\tilde{x}_1}[\tilde{x}_1] = 0$.)

An agent has a four times differentiable, increasing and strictly concave von Neuman-Morganstern utility function $U(W)$, and initial wealth W_0. The random variables $\tilde{x}_1$ and $\tilde{x}_2$ are bounded and independent. The risk premium, ρ_i, for bearing the risk, of the single random variable $x_i (i = 1, 2)$, satisfies

$$E_{\tilde{x}_i}[U(W_0 + \tilde{x}_i + \rho_i)] = U(W_0 + E_{\tilde{x}_i}[\tilde{x}_i]).$$

The risk premium, ρ_{1+2}, for bearing the risk of the random variable $\tilde{x}_1 + \tilde{x}_2$, satisfies

$$E_{\tilde{x}_1} E_{\tilde{x}_2}[U(W_0 + \tilde{x}_1 + \tilde{x}_2 + \rho_{1+2})] = U(W_0 + E_{\tilde{x}_1}[\tilde{x}_1] + E_{\tilde{x}_2}[\tilde{x}_2]).$$

Adding risks reduces the risk premium if

$$\rho_{1+2} < \rho_1 + \rho_2.$$

If $\tilde{x}_2$ is a small gamble, its risk premium is proportional to the Arrow-Pratt measure of absolute risk aversion or $-U''(W_0)/U'(W_0)$. Treating $\tilde{x}_1$ as part of the agent's endowment define the indirect utility function $V(x_2)$ of increments to wealth x_2, which is also von Neuman-Morganstern and defined as

$$V(x_2) = E_{\tilde{x}_1}[U(W_0 + \tilde{x}_1 + x_2)].$$

The expected utility of the agent bearing the risk of $\tilde{x}_1 + \tilde{x}_2$ is now expressed as $E_{\tilde{x}_2} V(\tilde{x}_2)$.

The incremental risk premium for bearing the risk of $\tilde{x}_2$, given that $\tilde{x}_1$ is in one's endowment is given by the Arrow-Pratt measure for the utility function $V(\cdot)$. The condition for $\rho_{1+2} < \rho_1 + \rho_2$ is for $V(\cdot)$ to be less risk averse than $U(\cdot)$, or

$$-\frac{E_{\tilde{x}_1}[U''(W_0 + \tilde{x}_1)]}{E_{\tilde{x}_1}[U'(W_0 + x_1)]} < -\frac{U''(W_0)}{U'(W_0)}.$$

Given our assumption that $E_{x_1}[\tilde{x}_1] = 0$, a sufficient condition can be directly obtained from Jensen's inequality. A sufficient condition is for the function $-U''(w)$ to be concave and $U'(w)$ to be convex, or $U''''(\cdot) \geqq 0$ and $U'''(\cdot) \geqq 0$ (with one inequality strict) over the range of $W_0 + \tilde{x}_1 + \tilde{x}_2$. (Clearly, $U'''' \leqq 0$ and $U''' \leqq 0$ (with one inequality strict) is sufficient for the reverse condition).

The assumption that $E_{\tilde{x}_1}[\tilde{x}_1] = 0$ is invalid if $\tilde{x}_1$ is a gamble which the agent accepts voluntarily. It is necessary to have $E_{\tilde{x}_1}[\tilde{x}_1] > 0$. Adding a voluntarily chosen gamble $\tilde{x}_1$ will not only place a mean preserving spread onto intitial wealth, it will increase mean wealth. To take account of the effect of this higher mean wealth on risk aversion, we augment the sufficient conditions described above with the condition for decreasing absolute risk aversion. This provides sufficient conditions for the "fallacy of large numbers" to be correct, rather than a fallacy, (for proof that this is equivalent to decreasing risk aversion, see Pratt (1964, Theorem 5) or Kihlstrom, Romer and Williams (1981, Corollary 2)).

The condition for decreasing absolute risk aversion at a point W is

$$U'''(W) > \frac{U''(W)^2}{U'(W)} > 0,$$

thus a sufficient condition is that over the entire domain of W

$$U'''(W) > \frac{U''(W)^2}{U'(W)}.$$

Combined with $U'''' \geqq 0$, we have a sufficient condition for diversification by adding risks to reduce the risk premium. Stronger characterizations can be obtained from stronger assumptions about the random variables $\tilde{x}_1$ and $\tilde{x}_2$.

These conditions extend beyond the case of $N=2$, because if over the relevant domain $U'''(\cdot) \geqq 0$ then $V'''(\cdot) \geqq 0$ and if $U''''(\cdot) \geqq 0$ then $V''''(\cdot) \geqq 0$. Finally, straight forward extension of Pratt (1964, theorem 5) shows that if $-U''/U'$ is decreasing in the relevant domain, then $-V''/V'$ is decreasing as well. A third independent gamble will further reduce the risk premium: $\rho_{1+2+3} < \rho_{1+2} + \rho_3 < \rho_1 + \rho_2 + \rho_3$.

A few examples may help to illustrate what is at work. With constant absolute risk aversion ($-U''(W)/U'(W) = k$ for all W), increasing the mean initial wealth is of no consequence and in addition,

$$\frac{-V''}{V'} = k\frac{E_{\tilde{W}}[U'(\tilde{W})]}{E_{\tilde{W}}[U'(W)]} = k,$$

implying that diversification by adding independent risks is of no consequence, because there are no wealth levels of lower risk aversion over which to average. The quadratic utility function $U(W) = W - (b/2)W^2$ has $U'''(\cdot) = U''''(\cdot) = 0$, therefore adding a zero mean gamble $\tilde{x}_1$ does not influence the risk aversion toward the independent gamble $\tilde{x}_2$. However, a voluntarily chosen gamble $\tilde{x}_1$ will have a positive expectation, and quadradic utility implies increasing absolute risk aversion. The per gamble risk premium will *increase* amd diversification will "hurt".

A simple example of a utility function which satisfies the conditions for diversification by adding risks to be beneficial is $U(W) = 0{\cdot}05W^3 - 60W^2 + 50{,}000W - 4{,}450{,}000$ which is increasing and concave and has $U'''(\cdot) > 0$, $U''''(\cdot) = 0$, and decreasing absolute risk aversion over the domain $W \in [0, 400)$. Suppose initial wealth is 100, notice that $U(100) = 0$. The gamble

$$\tilde{x}_1: \begin{cases} \frac{1}{2} & +32{\cdot}1983 \\ \frac{1}{2} & -30 \end{cases}$$

will be just acceptable; that is $E_{\tilde{x}_1}[U(100 + \tilde{x}_1)] = 0$. If $\tilde{x}_2$ is an independent identically distributed gamble, we find $E_{\tilde{x}_1}E_{\tilde{x}_2}[U(100 + \tilde{x}_1 + \tilde{x}_2)] = 209{\cdot}6$. The risk aversion toward the second gamble is reduced by accepting the first.

In contrast to diversification by subdividing risk, the value of diversification by adding risks depends critically on the form of agent's utility function. Given the lack of observability of preferences in practice, this limits the testability of this result on diversification when there is no binding wealth constraint. The results in Section 3, where the value of diversification arises only from binding wealth constraints, provide strong and testable results.

5. COMPARISON WITH LELAND–PYLE (1977) RESULTS

Leland–Pyle (1977) (L–P hereafter) develops an interesting model of costly signalling by entrepreneurs selling shares to the public. In contrast to the ex-post information asymmetry analysed in this paper, they focus on an *ex-ante* information asymmetry, where entrepreneurs know more than investors. This gives rise to an adverse selection problem, because if entrepreneurs of different types cannot be distinguished, all must sell securities at the same price, and there would be a large supply of securities by entrepreneurs with worthless projects. The model allows the entrepreneur an endogenous choice of investing

in other assets or retaining equity in his project. L–P show that retained equity serves as a costly signal of the entrepreneur's information about value. It is costly, because in equilibrium a risk averse entrepreneur retains some "project specific risk" of his project which would be avoided under full information.

Some preliminary thoughts on a theory of financial intermediation are presented in L–P although no analysis is developed. They suggest that financial intermediaries might expend resources to observe entrepreneur's *ex-ante* information and use the information to offer to buy securities from entrepreneurs, offering improved risk sharing. Intermediaries might do this to capture cost savings compared with information collection by investors, or to solve underproduction of information problems analysed by Grossman–Stiglitz (1980) and Chan (1983). Although L–P do not mention diversification, they seem to suggest that the intermediary collects information about many entrepreneurs and then signals the *ex-ante* prospects of its portfolio using the same "retained equity" costly signal which entrepreneurs can use individually. Such intermediation will be viable only if the per-entrepreneur risk sharing cost of signalling by the intermediary is lower than the per-entrepreneur cost of direct signalling without an intermediary. This is analogous to the conditions for viable intermediation in the delegated monitoring model analysed above. It is interesting to investigate whether the types of diversification analysed in Section 4 facilitate intermediation here. In the process of doing this, we correct an error in L–P which was not criticial to their analysis of individual entrepreneur signalling, but which is central to our extending the analysis to diversification and intermediation.[6] We present results in the text, and sketch the analysis in the Appendix.

The formal L–P signalling model analyses an entrepreneur endowed with a project which has a mean return observed only by him. It is common knowledge that it has a normal distribution with known variance σ^2. The entrepreneur and all investors have exponential utility (constant absolute risk aversion). The entrepreneur's preferences are common knowledge. Traded securities are valued in the market using the Capital Asset Pricing Model and public information. This implies that known market wide risks are "priced", while "specific risks" (those uncorrelated with the market portfolio) are not priced. The market will bear specific risks at no risk premium. The entrepreneur signals by issuing unlimited liability (riskless) debt, and equity at market prices, and by trading in the "market portfolio", with signalling conditions enforced by retaining a non-trivial amount of equity and its associated specific risk. This signalling is costly because of imperfect risk sharing—the risk averse entrepreneur retains a large amount of specific risk which could be sold off to the market with no risk premium under full information.

We introduce financial intermediation and diversification into the L–P model by assuming that there are N entrepreneurs with projects whose returns are distributed independently and identically and are independent of the market portfolio. (The results extend to the case where projects are correlated with the market portfolio, but independent conditional on the observed market portfolio). In the appendix, we demonstrate that the results of Section 4 carry over to the L–P model.

Diversification by adding independent risks occurs if the intermediary is modeled as a single agent who like everyone else in the L–P model has exponential utility. As is suggested by the analysis in Section 4, such diversification has no effect, because with constant absolute risk aversion the risk aversion toward any gamble is not affected by the presence of any other independent gambles. This implies that an agent signalling a given project will choose to retain a given fraction of its equity in a signalling equilibrium, irrespective of other independent projects he must signal, and that the marginal impact on his expected utility is not influenced by other independent projects he must signal. This

implies that financial intermediation based on diversification by adding risks is not viable given the L–P model because the intermediary signalling costs will be just as high as an entrepreneur's.

Diversification by subdividing risks occurs if there are N bankers working in the intermediary who all observe the *ex-ante* information of N entrepreneurs, and signal by each retaining equity in the intermediary's portfolio. Focusing for simplicity on the case of bankers with identical utility functions and identical independent projects with mean μ, and variance σ^2, this implies that each banker in the intermediary retains a fraction $1/N$ of total equity retained by insiders, and an equal fraction of each project. Because all bankers can observe each other's information and actions, they face no group moral hazard problem. Because of their risk sharing, we show that each banker's signalling decision is equivalent to that of a single entrepreneur signalling a project with mean $N\mu/N = \mu$ and variance $(1/N)^2 N\sigma^2 = \sigma^2/N$. As a result, diversification by subdividing risks has the same effect on each banker's expected utility as reducing the known variance of specific risk of a single project signalled directly by a single entrepreneur. We show in the Appendix that this diversification improves the expected utility of the agents in the intermediary (expected utility is a decreasing function of variance), implying that diversified intermediation is potentially viable. Put another way, the intermediary's signalling costs are lower than an entrepreneur's, because the intermediary's costs are equivalent to the signalling costs of an entrepreneur with a smaller variance of specific risk. This analysis corrects the erronous Proposition III in L–P, which states that an entrepreneur's expected utility is an *increasing* function of variance, and would have implied that even diversification by subdividing risk was counterproductive.

The results of our delegated monitoring intermediation model are consistent with the extension of the L–P analysis to intermediation. In particular, if the *ex-ante* information about the N entrepreneurs who contract with the intermediary is observed by the N bankers who as a team are the intermediary (diversification by subdividing risks), then the "delegated signalling" costs approach zero. The implication of this is an intermediary with primarily debt (deposits) in its capital structure and very little outside equity.

6. CONCLUSION

Diversification within the financial intermediary is the key to understanding why there is a benefit from delegating monitoring to an intermediary which is not monitored by its depositors. The intuitive reason for the value of diversification is slightly different in the model with risk neutral agents from the one with risk averse agents. In the risk neutral model, diversification is important because it increases the probability that the intermediary has sufficient loan proceeds to repay a fixed debt claim to depositors; in the limit, this probability is one, and the probability of incurring necessary bankruptcy costs goes to zero. In the model with risk aversion, but no binding constraints on non-negative consumption, diversification increases the intermediary's risk tolerance toward each loan, allowing the risk bearing necessary for incentive pruposes to be less costly. The general importance of diversification in financial intermediary theories is demonstrated by the similar results obtained from our analysis of a Leland–Pyle signalling model of intermediation.

Financial intermediaries allow better contracts to be used and allow Pareto superior allocations. This provides a positive role for financial intermediaries. The delegated monitoring model predicts well-diversified financial intermediaries with a capital structure which is mainly debt (deposits), with despite this high leverage, a low probability of

default. These predictions are in line with reality for most intermediaries. In addition, the insight that intermediaries must bear certain risks for incentive purposes has an important implication for the regulatory controversy involving the desirability of allowing banks to hedge in interest rate futures markets. Because interest rate risk is freely observable, it ought to be shared optimally, and permitting banks to sell such risk in the futures market effectively allows them to do so. Because risk sharing within the intermediary is constrained by binding incentive compatibility constraints, there is a reason to allow the bank to hedge against these risks (although a possible alternative is for the bank to force borrowers to do this hedging).

Commerical banks and insurance companies are the most obvious applications of this model. Another interesting application of the diversification by subdividing model is conglomerate firms. To the extent that members of subsidiary divisions can monitor each others actions at low cost, the conglomerate can allow the managers of the divisions to share the risks which they as a group must bear for incentive purposes. In Diamond–Verrecchia (1982), it is argued that the risks which managers must bear for incentive purposes are the firm specific risks because these are not observable elsewhere. If the cost of within conglomerate monitoring is fixed, a possible implication of our results is that firms with high firm specific risk will be most likely to join together into conglomerates.

An interesting implication of the delegated monitoring model is that intermediary assets will be illiquid. This is because the intermediary is delegated the task of observing information about each loan which no one else but the entrepreneur/borrower observes. In one sense, such assets are totally illiquid, as the intermediary contracts to hold them and enforce the contract, rather than sell them. If the intermediary were to sell a loan and transfer the monitoring and enforcement to someone else, the acquirer would have to incur the monitoring costs again, duplicating the effort of the first intermediary. These costs would be in addition to any physical costs of transfering ownership. Adverse selection of which loan an intermediary chooses to sell could be another complication caused by the private information possessed by the intermediary. The centralization of monitoring each loan by a single intermediary will mean that there are not active markets for these assets. All of these phenomenon are related to the concept of illiquidity. The resulting illiquidity of assets leads to another reason why financial intermediaries might improve on the allocations provided by competitive exchange markets; see Diamond–Dybvig (1983), where asset illiquidity is simply a result of the specified production technology. An interesting extension of these two models would be a model of the liqudity implications of private information within an intermediary.

Many "markets" for information services induce the delegated private information production analysed in this paper. Further study of the implications of this arrangement should produce new insights into financial markets and institutions, and possibly other types of markets and organizations.

APPENDIX

To focus on the role of diversification in reducing the signalling costs of an intermediary below those of individual entrepreneurs, we compare signalling costs for $N = 1$ and $N = 2$. We view the intermediary as equivalent to an entrepreneur with 2 projects. We assume that the projects of entrepreneurs monitored by the intermediary are mutually independent and uncorrelated with the "market portfolio", the one other traded risky asset in the L–P model. We follow L–P and assume that projects are so small that we can neglect the effect of adding them to the market. This implies we can alternatively assume that

the projects are independent conditional on the market portfolio, because trade in the market allows optimal linear sharing of market risk, and L–P analyze only linear risk sharing.

Agents in the intermediary have identical known exponential utility of wealth functions, $U(W) = -e^{-bW}$ where $b > 0$. Project i has returns $\tilde{x}_i + \mu_i$, where $\tilde{x}_i$ has a normal distribution with zero mean and known variance $\sigma^2_{x_i}$, and μ_i is known by the entrepreneur and intermediary, but not investors in the market. Given information which will be available to the market, the project is valued at its expectation, discounted by the rate of interest, r (because project is uncorrelated with the market, or alternatively, because investors are risk neutral). The intermediary chooses to retain a fraction α_i of the equity in the ith project, selling the remainder to outside investors and also issuing unlimited liability (riskless) debt. Absent some sort of self-selection or signalling mechanism, there will be a severe adverse selection problem. L–P solve for a fully separating signalling equilibrium, with sorting based on the value of α_i.

Using notation similar to L–P, define:

α_i = the fraction of the i-th project retained by the intermediary.
$\mu_i(\alpha_i)$ = the market's valuation schedule, expressing the μ_i inferred on the basis of α_i selected.
$V_i(\alpha_i)$ = the total market value of a project i implied by the schedule $\mu_i(\alpha_i)$. $V_i(\alpha_i) = \mu_i(\alpha_i)/(1+r)$.
W_0 = the initial wealth of the intermediary.
$\tilde{M}$ = the random return on the market portfolio.
β = the fraction of the market portfolio held by the intermediary.
V_M = price of the market portfolio.
K_i = initial outlay required for the ith project.
D_i = current value of riskless debt issued against i-th project (promise to pay $D_i(1+r)$).
Y = current value of riskless debt issued on "personal account". (The distinction between D_i and Y is not used here; we present it this way to be consistent with L–P.
r = riskless rate of interest.
W_1 = final wealth of intermediary.
$\sigma^2_{w_1}$ = variance of final wealth.

Given the assumption of exponential utility, and normal distributions of the projects and market portfolio, the intermediary maximizes $E[\tilde{W}_i] - (b/2)\sigma^2_{w_i}$. The budget constraint is:

$$W_0 + \alpha_1 D_1 + \alpha_2 D_2 + (1-\alpha_1) V_1(\alpha_1) + (1-\alpha_2) V_2(\alpha_2) - K_1 - K_2 - \beta V_M - Y = 0. \tag{A1}$$

Final wealth is

$$\tilde{W}_1 = \alpha_1[\tilde{x}_1 + \mu_1 - (1+r)D_1] + \alpha_2[\tilde{x}_2 + \mu_2 - (1+r)D_2] + \beta\tilde{M} + (1+r)Y. \tag{A2}$$

Substituting (A2) into (A1),

$$\tilde{W}_1 = \alpha_1[\tilde{x}_1 + \mu_1 - \mu_1(\alpha_1)] + \alpha_2[\tilde{x}_2 + \mu_2 - \mu_2(\alpha_2)] + \beta\tilde{M} - (1+r)V_M) + (W_0 - K_1 - K_2)(1+r) + \mu_1(\alpha_1) + \mu_2(\alpha_2). \tag{A3}$$

The intermediary chooses α_1, α_2 and β to maximize $E[\tilde{W}_1] - (b/2)\sigma^2_{\tilde{W}_1}$. Noting that

$\tilde{x}_1$, $\tilde{x}_2$, and $\tilde{M}$ are independent, the optimal α_1^*, α_2^*, and β^* satisfy:

$$[\mu_1 - \mu_1(\alpha_1^*)] + (1-\alpha_1^*)\mu_{\alpha_1}(\alpha_1) - \alpha_1 b\sigma_{x_1}^2 = 0, \tag{A4}$$

$$[\mu_2 - \mu_2(\alpha_2^*)] + (1-\alpha_2^*)\mu_{\alpha_2}(\alpha_2) - \alpha_2 b\sigma_{x_2}^2 = 0, \tag{A5}$$

and

$$[E[\tilde{M}] - (1+r)V_M] - \beta b\sigma_M^2 = 0. \tag{A6}$$

In a separating signalling equilibrium $\mu_i(\alpha_i) = \mu_i$. Solving (A4) and (A5) given this constraint yields

$$(1-\alpha_i)\mu_{\alpha_i}(\alpha_i) = b\alpha_i\sigma_{x_i}^2 \quad \text{for } i = 1, 2. \tag{A7}$$

Solving the differential equation (A7) yields

$$\mu_i(\alpha_i) = -b\sigma_{x_i}^2[\log(1-\alpha_i) + \alpha_i] + (1+r)K_i, \tag{A8}$$

plus an arbitrary constant. The least cost solution not subject to unraveling, is shown in L–P to have the constant = 0, implying that the market value of the ith project is

$$V_i(\alpha_i) = \frac{1}{1+r}[-b\sigma_{x_i}^2[\log(1-\alpha_i) + \alpha_i]] + K_i. \tag{A9}$$

For simplicity we analyse the case of independent and identically distributed (i.i.d.) projects, where $\mu_1 = \mu_2 = \mu$ and $\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_x^2$.

Diversification by *subdividing* risks occurs with an intermediary which consists of two agents each with risk aversion b who each retain a fraction $\alpha/2$ of each of two projects. Because their decisions are separable, they each make a decision for each project which is equivalent to that of a single agent endowed with a single project with mean $\mu/2$ and variance $\sigma_x^2/4$. This, in turn, is equivalent to a project with mean $= \mu$ and variance $= \sigma_x^2/2$, because from (A8), if α_i solves $\mu_i = [\log(1-\alpha_i) + \alpha_i]\sigma_{x_i}^2$, it also solves $a\mu_i = [\log(1-\alpha_i) + \alpha_1] \cdot a\sigma_{x_i}^2$. We can therefore analyse the comparative static effect of diversification by sub-dividing risks on an intermediary's expected utility by analysing the effect of reducing the variance of specific risk of a single project $i = 1$, holding its mean constant (we suppress the subscript "i"). This is given by

$$\frac{dE[U(\tilde{W}_1)]}{d\sigma_x^2} = \frac{dE[\tilde{W}_1]}{d\sigma_x^2} - \frac{b}{2}\frac{d\sigma_{W_1}^2}{d\sigma_x^2}. \tag{A10}$$

Because $E[\tilde{x}_i] = 0$ and $\mu = \mu(\alpha)$, (A3) shows that $dE[\tilde{W}_1]/d\sigma_x^2 = 0$. Turning to the variance of final wealth, note that it is given by $\sigma_{W_1}^2 = \alpha^2\sigma_x^2 + \beta^2\sigma_M^2$, implying

$$\frac{d\sigma_{W_1}^2}{d\sigma_x^2} = 2\sigma_x^2\alpha\frac{d\alpha}{d\sigma_x^2} + \alpha^2\frac{d\sigma_x^2}{d\sigma_x^2} + \frac{d(\beta^2\sigma_M^2)}{d\sigma_x^2}. \tag{A11}$$

Inspecting (A4) and (A6), $d(\beta^2\sigma_M^2)/d\sigma_x^2 = 0$, and by definition $d\sigma_x^2/d\sigma_x^2 = 1$. In equilibrium $\mu(\alpha) = \mu$, so one can apply the implicit function theorem to (A8), and obtain

$$\frac{d\alpha}{d\sigma_x^2} = -\frac{d\mu(\alpha)/d\sigma_x^2}{d\mu(\alpha)/d\alpha} = \frac{(1-\alpha)[\log(1-\alpha) + \alpha]}{\alpha\sigma_x^2}.$$

Inserting this into (A10) and (A11), one obtains

$$\frac{dE[U(\tilde{W}_i)]}{d\sigma_x^2} = -b\left[(1-\alpha)[\log(1-\alpha) + \alpha] + \frac{\alpha^2}{2}\right] < 0.$$

This is negative because it is defined over $\alpha \in (0, 1)$, is zero at zero, and decreasing in α. This corrects Proposition III in L–P, where the final α^2 term was omitted, leading them to conclude that the sign of the entire expression was positive. The intuition behind the correct result is clear: signalling is costly because of inferior risk sharing, if there is very little risk, the cost is low (if $\sigma^2 = 0$, there is no risk for the entrepreneur to share, and no need to go public). Thus diversification by subdividing risks can serve as a basis for viable financial intermediation in a L–P setting. As N, the number of independent projects, and number of bankers within the intermediary, grows without bound, the per-project risk premium goes to zero, because the total variance of wealth, per banker in the intermediary (σ_x^2/N) goes to zero.

Diversification by *adding independent* risks is modelled by adding a second i.i.d. project to the intermediary's portfolio, while the intermediary consists of a single agent with constant risk aversion of b. Inspecting (A4) and (A5), one finds that no terms involving another independent project enter, thus $\alpha_1 = \alpha_2$ and both are equal to the level that would prevail if there were only one project. Therefore, adding additional i.i.d. projects is equivalent to adding i.i.d. lotteries, and given exponential utility the analysis in Section 4 shows that the risk premium per project is not influenced by the number of independent projects. Therefore, diversification by adding i.i.d. projects does not reduce signalling costs in the L–P model and cannot serve as a basis for viable financial intermediation. It would be interesing to extend the L–P model to a utility function which implies that this type of diversification has value.

First version received August 1982; *final version accepted December* 1983 (*Eds.*).

I am grateful to S. Bhattacharya, G. Connor, P. Dybvig, B. Grundy, O. Hart, B. Holmström, M. Machina, D. Kyle, D. Romer, S. Ross, J. Tobin and R. Verrecchia for helpful comments. An earlier version of this paper was part of my dissertation submitted to the Yale University Department of Economics.

NOTES

1. Note that this formulation is without loss of generality, and contract $\phi(z)$, which specifies a payment of goods to the entrepreneur by the lender and is not a non-pecuniary penalty, can be expressed as a net payment $\hat{z} = \phi(z) - z$.

2. There are two equivalent ways to model the monitoring cost. One, mentioned in the text, is for the financial intermediary to experience no disutility from monitoring and enforcement, but to spend K in resources. In this case, to avoid messy notation, re-normalize so that "one unit" is defined as the sum of the amount each project requires plus K, the amount spent on monitoring by the intermediary. Alternatively, one can assume that monitoring does not require resources, but that the intermediary experiences disutility from monitoring and has the linear utilty function of wealth $U(W, N) = W - NK$, where N is the number of entrepreneurs monitored. In this case, no renormalization is required.

3. For example, in the identically distributed case with 1 project, the delegation cost is

$$D_1 = E_{\tilde{g}_1}[H_1 - \tilde{g}_1 | g_1 \leqq H_1],$$

while with 2 projects the per project delegation cost is

$$D_2 = \tfrac{1}{2} E_{\tilde{g}_1} E_{\tilde{g}_2}[H_2 - \tilde{g}_1 - \tilde{g}_2 | g_1 + g_2 \leqq H_2].$$

We know that the minimum feasible value of $H_2 \leqq 2H_1$ because if $H_2 = 2H_1$, the expected return to depositors is at least $2R$. This implies $D \leqq D_1 - C$, where

$$C = P(\Omega) E[H_2 - \tilde{g}_1 - \tilde{g}_2 | \Omega]$$

and Ω is the event "$g_1 + g_2 \geqq H_1 \cdot 2$, and either $[g_1 \leqq H_1$ or $g_2 \leqq H_1]$". If $\tilde{g}_1$ and $\tilde{g}_2$ have continuous distributions, $P(\Omega) > 0$ unless they are perfectly correlated, implying $C > 0$.

4. See Diamond–Verrecchia (1982) and Holmström (1982) for implications of this observation for the theory of mangerial capital budgeting.

5. Some analysis is presented of risk aversion measured from a stochastic initial "wealth" position in Kihlstrom, Romer and Williams (1981), Machina (1982), and Ross (1981). These papers do not address the problem of adding risks.

6. The analysis of L–P presented here was stimulated by a referee's noting that this paper's results seemed opposed to those of Proposition III of L–P, and his conjecture that L–P Proposition III might be incorrect.

REFERENCES

CAMPBELL, T. (1979), "Optimal Investment Decisions and the Value of Confidentiality", *Journal of Financial and Quantitative Analysis*, **14**, 913–924.
CHAN, Y. (1983), "On the Positive Role of Financial Intermediation in Allocation of Venture Capital in a Market with Imperfect Information", *Journal of Finance*, **38**, 1543–1568.
DIAMOND, D. W. and DYBVIG, P. H. (1983), "Bank Runs, Deposit Insurance and Liquidity", *Journal of Political Economy*, **91**, 401–419.
DIAMOND, D. W. and VERRECCHIA, R. E. (1982), "Optimal Managerial Contracts and Equilibrium Security Prices", *Journal of Finance*, **37**, 275–287.
HARRIS, M. and RAVIV, A. (1979), "Optimal Incentive Contracts with Imperfect Information", *Journal of Economic Theory*, **20**, 231–259.
HOLMSTRÖM, B. (1979), "Moral Hazard and Observability", *Bell Journal of Economics*, **10**, 74–91.
HOLMSTRÖM, B. (1982), "Moral Hazard in Teams", *Bell Journal of Economics*, **13**, 324–340.
KIHLSTROM, R., ROMER, D. and WILLIAMS, S. (1981), "Risk Aversion with Random Initial Wealth", *Econometrica*, **49**, 911–920.
LELAND, H. and PYLE, D. (1977), "Informational Asymmetries, Financial Structure, and Financial Intermediation", *Journal of Finance*, **32**, 371–387.
MACHINA, M. (1983), "Temporal Risk and the Nature of Induced Preferences" (Working paper, University of California-San Diego, Department of Economics).
PRATT, J. (1964), "Risk Aversion in the Small and the Large", *Econometrica*, **69**, 122–136.
RAMAKRISHNAN, R. and THAKOR, A. (1983), "Information Reliability and a Theory of Financial Intermediation" (Working paper, Indiana University).
ROSS, S. A. (1976), "The Arbitrage Theory of Capital Asset Pricing", *Journal of Economic Theory*, **13**, 341–360.
ROSS, S. A. (1981), "Some Stronger Measures of Risk Aversion in the Small and the Large with Applications", *Econometrica*, **49**, 621–638.
SAMUELSON, P. (1963), "Risk and Uncertainty: A Fallacy of Large Numbers", *Scientia*.
SCHUMPETER, J. (1939) *Business Cycles* (New York: McGraw-Hill).
SHAVELL, S. (1979), "Risk Sharing and Incentives in the Principal and Agent Relationship", *Bell Journal of Economics*, **10**, 55–73.
SMITH, C. W. and WARNER, J. B. (1979), "On Financial Contracting: An Analysis of Bond Covenants", *Journal of Financial Economics*, **7**, 117–161.
TOWNSEND, R. M. (1979), "Optimal Contracts and Competitive Markets with Costly State Verification", *Journal of Economic Theory*, **21**, 1–29.