

Wharton

**Financial
Institutions
Center**

*Is There an Optimal Size for the
Financial Sector?*

by
**Anthony M. Santomero
John J. Seater**

98-35-B

THE WHARTON FINANCIAL INSTITUTIONS CENTER

The Wharton Financial Institutions Center provides a multi-disciplinary research approach to the problems and opportunities facing the financial services industry in its search for competitive excellence. The Center's research focuses on the issues related to managing risk at the firm level as well as ways to improve productivity and performance.

The Center fosters the development of a community of faculty, visiting scholars and Ph.D. candidates whose research interests complement and support the mission of the Center. The Center works closely with industry executives and practitioners to ensure that its research is informed by the operating realities and competitive demands facing industry participants as they pursue competitive excellence.

Copies of the working papers summarized here are available from the Center. If you would like to learn more about the Center or become a member of our research community, please let us know of your interest.

Anthony M. Santomero
Director

*The Working Paper Series is made possible by a generous
grant from the Alfred P. Sloan Foundation*

Is There An Optimal Size for the Financial Sector ¹

February 1999

Abstract: This paper derives the optimal size of the financial sector using a general equilibrium framework that is an extension of Holmstrom and Tirole's 1997 paper. We show that the financial sector has a unique optimal size relative to the size of the economy as a whole. Creating and maintaining this sector requires diversion of some physical capital from production of output to monitoring that production. However, the efficiency gain in output production brought about by monitoring warrants the diversion. It is also found that the optimal size of the financial sector is independent of the state of the economy and does not vary over the business cycle.

JEL Classifications : E44, E42, E51, G2

Keywords : financial sector, intermediation theory, financial institutions

¹Anthony M. Santomero is at the Wharton School of the University of Pennsylvania.

John J. Seater is at the Economics Department of North Carolina State University.

I. Introduction

The advent of the Euro is the latest phase in the financial integration that is sweeping across Europe. Earlier events of special significance were the promulgation of the Second Banking Coordinating Directive, allowing banks to branch across national boundaries, and the establishment of the Financial Services Policy Group, designed to study inter-country issues arising from financial integration. It is clear that a unified continental financial services market is emerging in Europe. As that market develops, important questions will arise concerning the kind of market structure that will emerge, its appropriate size, and its organization. In many ways, both the developments and questions concerning them parallel those that have arisen in the United States over the last two decades with the increasing degree of financial integration taking place there. In both Europe and the United States, there are related questions concerning the public policies that should be enacted to guarantee that the resulting financial services industry is socially optimal - policies concerning mergers, types of services that can be offered by various types of institutions, capital adequacy requirements, and so on.

In this paper, we address a theoretical question that is important both for the positive and normative analysis of the financial industry, namely, what is the optimal size of that industry? This seems an obvious question for policy analysis, which concerns intervention in the financial industry precisely to guarantee some sort of social optimality, but the question also is important for a positive analysis, for determining the optimal size of the industry is closely related to analyzing the size that will emerge in competitive equilibrium. Thus the subject of this paper would seem important to several groups, including students of the financial industry, that industry's regulators, and both macroeconomists and macroeconomic policy makers. However, it

is only recently that economic theory has begun to address this important issue. This is, in large part, due to the fact that the financial sector has occupied a rather secondary position in formal macroeconomic theory for most of the past few decades. In Patinkin's (1965) neo-classical framework, the financial sector was limited to the demands and supplies of money and bonds; financial institutions played no significant role. Subsequent developments, such as Brunner and Meltzer (1968) and Tobin (1969), continued to assign to financial institutions only a minor role in determining macroeconomic equilibrium.

This view began to change with Bernanke's (1980) evidence that the Great Depression was at least partly the result of a reduction in the banking sector's ability to perform its evaluation and monitoring role. Bernanke's subsequent work with Gertler (1988, 1989) showed the importance of introducing into macroeconomic analysis the insights of the growing banking and intermediation literature. In particular, the work of Leland and Pyle (1977), Diamond (1984), and others had clearly established the importance of the monitoring function undertaken by such institutions.¹ Since this earlier work, a number of articles have developed a macroeconomic role for banks, emphasizing the value added by banks and often spotlighting banks' possible role in exacerbating business cycles and credit crunches.

Once one accepts the notion that the financial sector is important for real economic activity, however, some obvious questions come to mind. What is the appropriate or optimal size of the financial sector? What does that size depend on? How does it respond to changes in economic conditions? How do departures from the optimal size affect the economy? Holmstrom and Tirole's recent 1997 contribution is the first step in addressing some of these questions with

¹See Battacharya and Thakor (1992) and Allen and Santomero (1997) for reviews of this literature.

their analysis of the appropriate allocation of capital in a competitive market. Their results are provocative but are limited by their partial equilibrium setting.

In this paper, we develop a model of the economy similar in spirit to the Holmstrom and Tirole framework but in which all results are obtained in a general equilibrium framework, allowing us to address the interaction of the financial and real sectors more completely than has been done heretofore. We show that there is an optimal size for the financial sector, and that depends on some characteristics of the production and monitoring technologies. Interestingly, the optimal size of the financial sector is unrelated to the economic cycle and is not causally linked to things such as credit cycles, a result that contrasts sharply with those emerging from the partial equilibrium models of Holmstrom and Tirole and of Bernanke and Gertler. The general equilibrium framework also permits us to obtain other new results. For example, the size of the financial sector affects not only the level of output but also its growth rate. Also, both the magnitude and, more interestingly, even the direction of the response of aggregate consumption to a change in the financial sector's size depends on the current size of the financial sector relative to that of the economy as a whole and on several parameters of various behavioral functions. Finally, our results have implications for the regulation of financial intermediaries' capital ratios.

Section II of the paper presents the background for our approach. Section III builds a simple static model to establish some fundamentals and lay the foundation for the dynamic model. Section IV presents the dynamic model. Section V concludes the paper.

II. The Microfoundations of the Financial Intermediation Model.

Our approach is motivated by Holmstrom and Tirole's (1997) analysis of financial

intermediation. We therefore begin by summarizing the basic view of the world captured in their model to provide the foundation for our own analysis.

1. *Ownership of Capital.* There are three kinds of agents: firms, intermediaries, and uninformed investors (hereafter called households for simplicity). Each type of agent holds capital. The individual firm holds an amount A of capital, and all firms together hold $K_f = \int A dG(A)$ where G is the distribution of A across firms. Intermediaries and households hold the total amounts of capital K_m , and K_h , respectively.

2. *Investment.* Entrepreneurs own firms that undertake investment projects, all of which are of the same fixed size I . The return to a project is random in that a project can either succeed or fail. It has a return of R if it succeeds, and a return of zero if it fails. The probability of success depends on how the firm behaves. Firms can behave diligently or can shirk. Firms derive a benefit of unspecified nature if they shirk; that benefit is denoted B . If the firm behaves diligently, the probability of success is p_h ; if the firm shirks, the probability is $p_l < p_h$. This is a similar set up to the return from effort modeled elsewhere by Allen and Gale (1988). Finally, there is an opportunity cost of undertaking an investment project equal to αI , where $\alpha > 0$ is the return the firm could get on its capital if it invested in the financial market instead of in its own project. The firm thus faces two possible expected net returns on its investment project:

$$\begin{array}{ll} p_h R - \alpha I & \text{if it is diligent} \\ p_l R - \alpha I + B & \text{if it shirks} \end{array}$$

3. *Borrowing and Lending: No Intermediaries.* It is assumed that shirking never is profitable if the firm finances its investment project entirely with its own funds:

$$p_l R - \alpha I + B < p_h R - \alpha I$$

This assumption is merely for convenience in the discussion below. The essential element of what follows is that firms that borrow some of their capital are more likely to shirk than firms that do no borrowing (intuitively, because the former have less at stake than the latter). The easiest way to frame the argument is simply to assume that firms that do no borrowing also do not shirk. Some firms, however, must borrow if they are going to invest. Those are the firms that do not own enough capital to undertake an investment project, that is, for whom $I > A$. Such firms borrow from households. If the investment project is successful, the firm pays its creditors the contracted part of the total return and keeps the rest for itself. If the project fails, the creditors are paid nothing. Thus the two expected returns facing the firm now are

$$\begin{array}{ll}
 p_h(R-P) - \alpha I & \text{if it is diligent} \\
 p_l(R-P) - \alpha I + B & \text{if it shirks}
 \end{array}$$

where P is the contracted payment to the creditors.

Clearly, once the firm has financed some of its project with borrowing, it has an increased incentive to shirk because its expected return from the project itself is lowered by the required payments to its creditors. This kind of situation has been modeled extensively elsewhere in the finance literature as the incentive effect of debt. As that literature shows, the firm that seeks to borrow must guarantee its creditors that it will not shirk, which it does by paying itself a large enough fraction of the total expected return $p_h R$ to make shirking unprofitable. It then pays the creditors out of the residual return. Using this framework, Holmstrom and Tirole have shown that only firms with sufficiently large values of A can borrow. Small firms cannot borrow because they cannot pay themselves enough to guarantee that they will not shirk and simultaneously pay a competitive rate of return to their creditors.

4. *Borrowing and Lending with Intermediaries.* Next, consider an environment in which intermediaries lend to firms. Intermediaries finance these loans with their own capital K_m and by borrowing from households. Intermediaries also monitor the firms to which they lend. This monitoring reduces the benefit of shirking to $b < B$ but costs the intermediary C per firm monitored. A firm that is too small to borrow directly from households may be large enough to borrow from the intermediary. To get a loan from an intermediary, the firm must agree to be monitored, and it must pay the intermediary a premium to cover the costs of intermediation.² Because of this premium, borrowing from the intermediary is more costly than borrowing directly from households, but it will be worthwhile if b is sufficiently less than B .

It is assumed that the various parameters satisfy the conditions necessary for intermediary lending to occur. Financial intermediaries then lend only to firms of intermediate size. Large firms either do not borrow at all or borrow directly from households, because the absence of the monitoring cost premium makes it cheaper to do so; small firms still do not have enough capital to guarantee that they will not shirk. Thus there is a range of firms (A_l, A_u) interior to the support of the distribution $G(A)$ that receives loans from the financial intermediaries.³ The bounds A_l and A_u both depend positively on the market rate of return α , and A_l also depends positively on the expected gross return to intermediary capital (the expected gross payment $p_h R$ less the monitoring cost C divided by the amount of intermediary capital K_m). Competition among intermediaries

² This set up is similar to the argument in Fama (1985).

³ Actually, Holmstrom and Tirole show that firms that borrow from intermediaries also borrow directly from households. By financing some of their loans through intermediaries, the firms reduce the exposure of the households in two ways. First, the firms are monitored, which reduces their return from shirking. Second, the total amount of capital that the households lend to the firms is reduced by the amount the firms obtain from the intermediary.

forces them to invest their own capital K_m in the firms. Doing so regulates the rate of return earned by intermediaries in such a way as to make the market for capital clear.

5. *Some Important Results.* Three types of capital tightening are possible in this model: a collateral squeeze, a credit crunch, and a savings squeeze, in which K_f , K_m , and K_h fall, respectively. In all three cases, aggregate investment falls, and A_1 rises. Consequently small, poorly capitalized firms lose their financing in any of these situations. In an extension of the model to the case where investment size I is not fixed but can be chosen by firms, Holmstrom and Tirole show that the “solvency ratios” $r_f = K_f/(K_f+K_m+K_h)$ and $r_m = K_m/(K_m+K_h)$ respond to the three kinds of credit tightening in different ways. In a collateral squeeze, r_f falls, and r_m rises. In a credit crunch, exactly the opposite occurs: r_f rises, and r_m falls. In a savings squeeze, both r_f and r_m rise. These last results suggest that optimal regulation of financial institutions’ capital ratios may have to allow for cyclical variation in the minimum required ratios. However, the allowance will depend on the source of the cycle. Two types of reductions in capital availability lead to increases in the optimal value of r_m ; the remaining type leads to a decrease.

6. *Some Limitations.* The Holmstrom-Tirole model is very interesting and offers many insights into the behavior of the credit market and its interaction with the real sector. It is limited, however, to a partial equilibrium analysis of the credit market. In their model, the quantities of firm capital K_f and intermediary capital K_m are fixed and do not respond to economic conditions, and the source of household capital is unspecified. In reality, households own all the capital, K_f and K_m as well as K_h . Thus, changes in one type of capital presumably would come at least in part from opposite changes in one or both of the other types. Also, total capital can change only if total output changes or if households alter their consumption. The Holmstrom-Tirole model

ignores the household sector's optimization problem entirely. Finally, the Holmstrom-Tirole model leaves unexplained the reasons for the three types of credit tightening. Why should intermediary capital (or either of the other types) change? Shouldn't the reason have implications for the other kinds of capital? It is unclear how inclusion of these various aspects of the aggregate economy would alter the conclusions of the model. We therefore examine a version of the model in a general equilibrium setting.

III. A Static Model

Like Holmstrom-Tirole, we assume the only function that intermediaries perform is the investigation and/or monitoring of firms. The banking literature referenced above does concentrate on this role as unique to the intermediary sector. According to that literature, other intermediary products (such as conversion of small loans into large loans or conversion of short term loans into long term loans,) can be seen as by-products or at least joint products of monitoring. While households can perform many of the functions of a bank on their own, it is our view that one key reason that households use the bank is to collect information instead of collecting it themselves. Banks are specialized in performing precisely this function. In any case, here, we will restrict attention to intermediaries as investigators and monitors of firms.

Given this simplification, we must find a tractable way to represent the provision of monitoring services in the context of the aggregate economy, which turns out to be the major difficulty in constructing the general equilibrium model. Once that has been done, we can introduce a straightforward household utility function and obtain the general equilibrium solution for the economy, which also is quite straightforward. We begin with a static model; the results

obtained from it then carry over to a dynamic model that we discuss later.

1. *Basic Production.* The underlying production technology is the AK production function:

$$(1) \quad Y = AK$$

The AK technology has been widely used in the growth literature as the simplest production function permitting endogenous growth. Several more sophisticated models of technical progress end up with equilibrium solutions that are merely elaborate versions of the AK model.⁴ We simplify by just assuming AK production at the outset.

There is a continuum of firms distributed uniformly from 0 to F_U . We assume the distribution of firms is fixed, so there is no variation in the number of firms or in the concentration of mass along the interval $[0, F_U]$. Firms differ in size, measured by the firm's capital stock. The firm's capital is proportional to its position in the interval $[0, F_U]$; firm F has capital stock κF , where κ is the factor of proportionality, constant across firms. The distribution of capital thus is also uniform, with the largest capital stock being $K_U = \kappa F_U$. We do not address in detail why a distribution of firms exists at all. One obvious possibility is that the variance of returns differs by firm size. Perhaps small firms are innovators and so have a higher variance of returns than larger, established firms. A formal analysis of such a possibility requires making each firm's return random and also requires one to examine household (i.e., investor) behavior toward risk. Such issues are well beyond the scope of the present paper, although they would provide interesting grounds for extensions of the present analysis. We simply assume the existence of the appropriate distribution of firm sizes.

⁴See, for example, the discussions of the variety and quality ladder models in Barro and Sala-i-Martin (1995, chapters 6 and 7).

The aggregate capital stock is

$$\begin{aligned}
 (2) \quad K^* &= \int_0^{F_U} K(F) dF = \int_0^{F_U} \kappa F dF = \int_0^{K_U} K \kappa^{-1} dK \\
 &= \frac{1}{2\kappa} K_U^2
 \end{aligned}$$

It may seem that the aggregate capital stock is not proportional in the individual firms' stocks, but it is. If we change every firm's capital by the same proportion so that the new factor of proportionality is $\kappa' = p\kappa$, then the aggregate capital stock is

$$\begin{aligned}
 (3) \quad K^* &= \int_0^{F_U} K(F) dF = \int_0^{F_U} p\kappa F dF = \int_0^{pK_U} K(p\kappa)^{-1} dK \\
 &= p \frac{1}{2\kappa} K_U^2
 \end{aligned}$$

The proportionality parameter κ plays no role in the subsequent analysis, so henceforth we assume $\kappa=1$ for simplicity. Thus we can write aggregate capital in terms of the fixed distribution of firm capital as

$$\begin{aligned}
 (4) \quad K^* &= \int_0^{K_U} K dK \\
 &= \frac{1}{2} K_U^2
 \end{aligned}$$

From the preceding analysis we know that proportional changes in all firms' capital simply multiplies this integral by the relevant factor of proportionality. This fact will be useful in our analysis of the growth path of the economy.

Similarly, aggregate output is

$$(5) \quad Y^* = \int_0^{K_U} AK dk = A \frac{K_U^2}{2}$$

$$= AK^*$$

which, like aggregate capital, responds proportionally to a given proportional change in all firms' capital stocks. All variation in output arises from changes in the amount of capital firms own. To be consistent with our assumptions on the distributions of firms and capital, the only variations in capital that we permit are equiproportional changes in all firms' capital. As we have seen above, the aggregate capital stock changes by the same proportion, so (5) tells us that aggregate output changes by that proportion, too.

2. Inefficient Production. Firms may behave inefficiently, perhaps because managers receive some private benefit such as excessive perks from inefficient behavior. Inefficient behavior leads to reduced production:

$$(6) \quad Y = vAK$$

where $0 \leq v \leq 1$. The probability that a firm behaves inefficiently is w . Thus the expected output of a firm is

$$(7) \quad Y = wvAK + (1-w)AK$$

$$= AK[1 - w(1-v)]$$

We assume that the inefficiency probability w is inversely related to firm size; in particular, we assume the linear relation

$$(8) \quad w = w_0 - w_1 K$$

$$= 1 - K/K_U$$

Given this function, the smallest firm ($K=0$) is guaranteed to be inefficient ($w=1$), and the largest

($K=K_U$) is guaranteed to be efficient ($w=0$). The assumed inverse relation between w and firm size is justifiable if the benefit of being inefficient is unrelated to the size of the firm. For example, the benefit could be extra leisure obtained by shirking responsibilities (think of an efficiency wage framework). The cost of inefficiency, however, is the opportunity cost of foregone output $(1-v)AK$, which falls with firm size K . Thus inefficiency is more likely for small firms.

Combining (7) and (8) gives the expected output for a firm of size K :

$$(9) \quad Y = vAK + [A(1-v)/K_U]K^2$$

3. *Production with Monitoring.* Monitoring of firms by financial intermediaries increases the expected output of the firm. We can think of the mechanism as either a reduction in the probability w or an increase in the inefficiency parameter v , that is, a reduction in the cost of inefficiency. Under the first mechanism, the monitored firm is less likely to be inefficient but, if it does act inefficiently, it is just as inefficient as if it had not been monitored. Under the second mechanism, the firm is just as likely to be inefficient as if it were not monitored but its departure from efficiency is less. In reality, both mechanisms probably function, but we assume just one for simplicity. The two turn out to have virtually identical implications, so we choose the second mechanism for concreteness.

The expected output of the monitored firm is

$$(10) \quad \begin{aligned} Y &= AK[1-w(1-v_m)] \\ &= mvAK + [A(1-mv)/K_U]K^2 \end{aligned}$$

where the monitoring effectiveness parameter m satisfies $1 < m \leq v^{-1}$. The benefit of monitoring a firm is the increase in output obtained:

$$\begin{aligned}
(11) \quad Y|_{\text{with monitoring}} - Y|_{\text{without monitoring}} &= AK[1-w(1-vm)] - AK[1-w(1-v)] \\
&= (m-1)v wAK \\
&= (m-1)vA(K-K^2/K_U)
\end{aligned}$$

which is quadratic in K with zeroes at 0 and K_U and a maximum of $(m-1)vAK_U/4$ at $K_U/2$.

Monitoring for a given firm is a zero-one decision: either the firm is monitored or it is not. We do not include here the possibility of changing the intensity of monitoring a given firm. Thus the only decision concerning monitoring is the choice of which firms to monitor. This decision depends on the nature of the monitoring cost, which we treat as entirely an opportunity cost. Monitoring is achieved by diverting capital from production to monitoring. We assume that all firms contribute equiproportionally to K_m , so that $K_m = \mu K$ with $0 \leq \mu \leq 1$ and the aggregate stock of monitoring capital is just $K_m^* = \mu K^*$. This assumption can be motivated by supposing that a social planner chooses the level of monitoring capital and finances it with a proportional wealth tax or by assuming that households divide their assets between manufacturing firms and financial intermediaries. We discuss the social planner in more detail shortly. The upshot is that the social monitoring cost is the foregone output due to diverting capital from production, equal to $AK_m^* = \mu AK^*$.

We also suppose that the amount of effort required to monitor a firm is proportional to the firm's size. This is equivalent to assuming that it takes a fixed amount of effort to monitor one unit of capital. Firms with more capital then require proportionately more monitoring effort. In the aggregate, then, the fraction ϕ of total capital that is monitored is proportional to the amount of monitoring capital K_m^* :

$$\begin{aligned}
(12) \quad \phi &= \phi_0 \frac{K_m^*}{K^*} \\
&= \phi_0 \mu
\end{aligned}$$

The fraction ϕ is measuring the efficiency of monitoring, that is, the amount of productive capital that is monitored by a given amount of monitoring capital. In contrast, the effectiveness parameter m in equation (10) measures the impact of monitoring on the performance of capital that is monitored. It seems reasonable to suppose that it takes much less than the total capital stock to achieve monitoring of all productive capital, so we suppose that ϕ_0 is much greater than 1. Also, it is impossible for ϕ to exceed 1, so optimal μ must satisfy $0 \leq \mu \leq \phi_0^{-1}$.

4. *Optimal Monitoring.* We are interested in the socially optimal amount of monitoring, so we suppose there is a social planner who makes all allocation decisions for the economy. The planner seeks to maximize social welfare, which is equivalent to maximizing the utility of the representative household. To avoid unnecessary complications, we assume all households are alike, so the representative household is the same as any actual household. We also suppose households have the Constant Relative Risk Aversion utility function

$$(13) \quad u(C_t) = \frac{C_t^{1-\theta} - 1}{1-\theta}$$

where C is consumption per person. The planner seeks to maximize (13) subject to the aggregate resource constraint $Y_t^* = C_t^*$, where C^* is aggregate consumption. In this static, one-period setting, maximizing utility is equivalent to maximizing current output. The only instrument available to the planner for affecting current output is monitoring, so he chooses monitoring to maximize Y_t^* . It will turn out that the choice he makes in this static setting will be the same

choice he makes in a dynamic, multi-period setting, which we discuss later.

There are two aspects to the choice of monitoring: how much total monitoring to do, which is determined by the choice of μ (or, equivalently, of ϕ or K_m^*), and which capital to monitor. For any given total amount of monitoring, the choice of which capital to monitor is straightforward. From the continuity of all functions, it is clear that this capital will fall in a continuous interval, which is easily represented. The total amount of capital that is monitored is $\phi K^* = \phi_0 \mu K^*$. The continuous interval of monitored capital then can be written as

$$(14) \quad [K^C - \frac{\phi_0 \mu K^*}{2}, K^C + \frac{\phi_0 \mu K^*}{2}] = [K^C - \frac{\phi_0 \mu K_U^2}{4}, K^C + \frac{\phi_0 \mu K_U^2}{4}]$$

where K^C is the midpoint of the interval. Making an optimal choice of monitoring consists of choosing K^C and μ .

Let M be an indicator of monitoring, equal to 1 if a firm is not monitored and m if it is monitored. Then aggregate output with inefficiency and monitoring is

$$(15) \quad \begin{aligned} Y^* &= \int_0^{K_U} (1-\mu)AK[1-w(1-\nu m)]dK \\ &= \int_0^{K_U} (1-\mu)AK[1-w(1-\nu)]dK + \int_{K^C - \phi_0 \mu K_U^2/4}^{K^C + \phi_0 \mu K_U^2/4} (1-\mu)AKw\nu(m-1)dK \\ &= \int_0^{K_U} AK[1-w(1-\nu)]dK - \int_0^{K_U} \mu AK[1-w(1-\nu)]dK \\ &\quad + \int_{K^C - \phi_0 \mu K_U^2/4}^{K^C + \phi_0 \mu K_U^2/4} (1-\mu)AKw\nu(m-1)dK \end{aligned}$$

In the last line, the first term is aggregate output with no monitoring, the second term is the

opportunity cost of monitoring (the output lost by diverting $K_m^* = \mu K^*$ capital from production to monitoring), and the third term is the output gained through the efficiency increases due to monitoring. Only the last term depends on K^C , so K^C is chosen to maximize it. Substituting $w = (1 - K/K_U)$ gives the following expression for the last term in (15):

$$(16) \quad (1 - \mu)v(m-1)AK \int_{K^C - \phi_0 \mu K_U^2/4}^{K^C + \phi_0 \mu K_U^2/4} (K - K^2/K_U) dK$$

This expression is maximized for $K^C = K_U/2$, the midpoint of the range of K .⁵

Substituting this value of K^C into (15) and carrying out the integration gives the following expression for aggregate output as a function of μ :

$$(17) \quad \begin{aligned} Y^* &= (1 - \mu)A \frac{K_U^2}{2} \left[\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 \mu \left(1 - \frac{1}{12} \phi_0^2 \mu^2 \right) \right] \\ &= (1 - \mu)A K^* \left[\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 \mu \left(1 - \frac{1}{12} \phi_0^2 \mu^2 \right) \right] \end{aligned}$$

The entire choice of optimal monitoring thus reduces to choosing μ .

The first-order condition for μ is

$$(18) \quad \begin{aligned} 0 &= -\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 - \frac{v(m-1)}{2} \phi_0 \mu \\ &\quad - \frac{v(m-1)}{16} \phi_0^3 \mu^2 + \frac{v(m-1)}{12} \phi_0^3 \mu^3 \end{aligned}$$

This expression is a cubic in μ and gives little immediate insight into the optimal value of μ .

However, by rewriting (18), we can learn something about the optimal μ . The production

⁵Notice that this is the midpoint of the range of capital, not of firms. Denote by F^C the firm that holds the capital stock K^C . As much capital above K^C is monitored as below it. Consequently, the monitored firms larger than F^C fewer in number than the monitored firms smaller than F^C .

function can be written as

$$(19) \quad Y^* = (1-\mu)AK^*g(\mu)$$

where

$$(20) \quad \begin{aligned} g(\mu) &= \frac{\nu+2}{3} + \frac{\nu(m-1)}{4} \phi_0 \mu \left(1 - \frac{1}{12} \phi_0^2 \mu^2\right) \\ &= \frac{\nu+2}{3} + \frac{\nu(m-1)}{4} \left(\phi_0 \mu - \frac{1}{12} \phi_0^3 \mu^3\right) \end{aligned}$$

If we define $G(\mu) \equiv (1-\mu)g'(\mu)$, then the first-order condition can be written as

$$(21) \quad \begin{aligned} (1-\mu)g'(\mu) &= g(\mu) \\ G(\mu) &= \end{aligned}$$

The function g is cubic and so has three possible roots. To find them, we begin by noting that the first and second derivatives of g are

$$(22) \quad \begin{aligned} g'(\mu) &= \frac{\nu(m-1)}{4} \left(\phi_0 - \frac{1}{4} \phi_0^3 \mu^2\right) \\ g''(\mu) &= -\frac{\nu(m-1)}{2} \phi_0^3 \mu \end{aligned}$$

If we evaluate g , g' , and g'' at 0, we obtain the following results:

$$(23) \quad \begin{aligned} g(0) &= \frac{\nu+2}{3} \in \left[\frac{2}{3}, 1\right] \quad \text{because } \nu \in [0, 1] \\ g'(0) &= \frac{\nu(m-1)}{4} \phi_0 > 0 \\ g''(0) &= 0 \end{aligned}$$

Thus 0 is an inflection point of g , and g is positive and rising there. These results mean that g has the general shape shown in Figure 1, with two negative roots and one positive root. We have drawn Figure 1 with distinct negative roots, but they could be a double root or two imaginary roots. We do not dwell on this detail because only the positive root is economically meaningful.

To find the maximum of g in the positive quadrant, we set $g'(0)$ equal to zero and solve for μ , obtaining $\mu = \pm 2\phi_0^{-1}$. Only the positive value is economically meaningful, so $g(\mu)$ has single maximum in the positive quadrant at $\mu = 2\phi_0^{-1}$; the value of g at that point is $(vm+2)/3$. Recall that μ is restricted to lie in the interval $[0, \phi_0^{-1}]$, so $g(\mu)$ is rising over the entire permissible range of μ .

The function $g'(\mu)$ is positive and falling over all permissible values of μ , reaching 0 at the value $\mu = \phi_0^{-1}$. The function $G(\mu)$ therefore also has these characteristics because $(1-\mu)$ is positive for all values of μ in the interval $[0, \phi_0^{-1}]$. We have drawn G in Figure 2.

The optimal value of μ , denoted μ^* , occurs at the intersection of the $g(\mu)$ and $G(\mu)$ functions. The existence condition for a positive value of μ^* is

$$(24) \quad \frac{v(m-1)}{4}\phi_0 > \frac{v+2}{3}$$

If this inequality is satisfied, the vertical intercept of $G(\mu)$ is above that of $g(\mu)$, and the two functions intersect at a positive value of μ . If (24) is not satisfied, the costs of monitoring exceed the benefits, and no there will be no monitoring (i.e., $\mu^* = 0$). If (24) is satisfied, an intersection may occur at a value of μ above ϕ_0^{-1} , which is the upper bound for μ . In that case, μ^* would equal ϕ_0^{-1} itself, implying that the fraction $\phi = \phi_0\mu$ of productive capital that is monitored is 1. In any case, there is only one solution for μ^* , even though the first-order condition is cubic in μ . The solution for μ is illustrated in Figure 3, which is drawn for the intermediate case where $0 < \mu^* < \phi_0^{-1}$.

The optimal value μ^* is the goal of our quest. It determines all aspects of socially optimal monitoring, including the appropriate quantity of society's capital that should be used for

monitoring, that is, the optimal size of the financial sector. It would be of interest to investigate how the sector's size is related to total capital, the degree of inefficient behavior, and the impact of monitoring on output. It is surprisingly difficult to say much about how μ^* responds to changes in the three parameters ϕ_0 , v , and m . Total differentiation of the first-order condition yields an expression of the form $d\mu = X_1 d\phi_0 + X_2 dv + X_3 dm$, but the X_i coefficients are highly non-linear in the parameters and generally of ambiguous sign. In the next section, we examine how μ^* responds to changes in the state of the economy.

Before we move to the dynamic model, we should address one issue concerning the distribution of firms. We have assumed implicitly that the distribution of firms is invariant to the existence or scope of monitoring. This assumption is unlikely to be literally correct in practice. Monitoring is applied only to middle-sized firms, so it raises their productivity relative to all other firms. In response, one would expect the social planner to shift resources to middle-sized firms away from the large and small firms. We have ignored this possibility. We doubt that any of our conclusions would be affected by allowing the distribution to change in response to the existence of monitoring, at least as long as the distribution did not degenerate to a point located at the mean size of firms. It seems likely that degeneration would not occur for the same reasons that the distribution exists in the first place. Monitoring alters the relative returns to firms of various sizes, but it does not eliminate whatever differences across firms leads to a non-degenerate distribution in the absence of monitoring. Thus we expect that the general character of the distribution of firms would be the same with and without monitoring. In that case, our analysis can be regarded as an approximation that ignores the changes in the distribution that monitoring might induce.

IV. A Dynamic Model

We turn now to a dynamic model in which a social planner chooses a path of μ to maximize lifetime utility of a representative household.

1. *The Planner's Problem.* Population growth plays no important role in this model, so we set it to zero. The planner seeks to maximize the present value of the representative household's lifetime utility

$$(25) \quad \int_0^{\infty} \frac{C_t^{1-\theta} - 1}{1-\theta} e^{-\rho t} dt$$

subject to the constraint on capital growth

$$(26) \quad \begin{aligned} dK^*/dt &= Y_t^* - C_t - \delta K_t^* \\ &= (1-\mu_t)AK_t^* \left[\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 \mu_t \left(1 - \frac{1}{12} \phi_0^2 \mu_t^2 \right) \right] - C_t - \delta K_t^* \\ &= AB(\mu_t)K_t^* - C_t - \delta K_t^* \end{aligned}$$

where

$$(27) \quad B(\mu_t) \equiv (1-\mu_t) \left[\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 \mu_t \left(1 - \frac{1}{12} \phi_0^2 \mu_t^2 \right) \right]$$

is the benefit of monitoring. The Hamiltonian for the social planner's problem is

$$(28) \quad H(C_t, \mu_t) = \frac{C_t^{1-\theta} - 1}{1-\theta} e^{-\rho t} + \psi_t [AB(\mu_t)K_t^* - C_t - \delta K_t^*]$$

The necessary conditions are (26) and

$$(29) \quad d\psi/dt = -\frac{\partial H}{\partial K^*} = -\psi[AB(\mu_t) - \delta]$$

$$(30) \quad \begin{aligned} \partial H / \partial C &= C_t^{-\theta} e^{-\rho t} - \psi_t \\ &= 0 \end{aligned}$$

$$\begin{aligned}
(31) \quad \partial H / \partial \mu &= \psi A K_t^* \left[-\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 - \frac{v(m-1)}{2} \phi_0 \mu_t \right. \\
&\quad \left. - \frac{v(m-1)}{16} \phi_0^3 \mu_t^2 + \frac{v(m-1)}{12} \phi_0^3 \mu_t^3 \right] \\
&= 0
\end{aligned}$$

$$(32) \quad K_0^* \text{ given}$$

$$(33) \quad \lim_{t \rightarrow \infty} K_t^* \psi_t = 0$$

We already can say something important about the optimal path of μ . Notice that (31), the first-order condition for μ , reduces to

$$\begin{aligned}
(34) \quad 0 &= -\frac{v+2}{3} + \frac{v(m-1)}{4} \phi_0 - \frac{v(m-1)}{2} \phi_0 \mu_t \\
&\quad - \frac{v(m-1)}{16} \phi_0^3 \mu_t^2 + \frac{v(m-1)}{12} \phi_0^3 \mu_t^3
\end{aligned}$$

which is exactly the same as the first-order condition (18) from the static model. Consequently, the optimal value of μ_t is equal to its value μ^* in the static model and depends only on the three parameters ϕ_0 , v , and m . In particular, it does not depend on the state of the economy (K_t^* , ψ_t) or the path of consumption C_t . There are two important implications of this result. First, μ^* is independent of the size of the economy, which in turn means that the stock of monitoring capital K_m^* is just proportional to the aggregate stock of capital K^* . The relative size of the monitoring sector is constant, so the absolute size is proportional to the size of the economy as a whole. Second, μ^* displays no cyclical behavior, in contrast to the Holmstrom-Tirole model. In this model, business cycles would be induced by shocks to A ; such shocks change the paths of C and K^* , as we show momentarily. However, A , C , and K all are absent from (34), so μ^* does not

depend on them.

2. *Growth Rates.* Our aggregate model is an extended form of the standard AK model from growth theory with A replaced by $AB(\mu)$. We therefore obtain the growth rates in the usual way.

Differentiating the first-order condition (30) for consumption with respect to time and rearranging gives the growth rate of consumption:

$$(35) \quad \begin{aligned} \gamma_C &\equiv (dC/dt)/C \\ &= \frac{1}{\theta}[AB(\mu) - \delta - \rho] \end{aligned}$$

which implies a time path for consumption of

$$(36) \quad C_t = C_0 e^{\frac{1}{\theta}[AB(\mu) - \delta - \rho]t}$$

We are interested in the case where $\gamma_C > 0$, so we suppose that $AB(\mu) > \delta + \rho$. We also want to ensure bounded lifetime utility. Lifetime utility along the optimal path is

$$(37) \quad \begin{aligned} &\int_0^{\infty} \frac{1}{1-\theta} e^{-\rho t} (C_0 e^{\frac{1-\theta}{\theta}[AB(\mu) - \delta - \rho]t} - 1) dt \\ &= \frac{1}{1-\theta} \int_0^{\infty} (C_0 e^{\frac{1-\theta}{\theta}[AB(\mu) - \delta - \rho - \rho]t} - e^{-\rho t}) dt \\ &= \frac{1}{1-\theta} \int_0^{\infty} [X_1(t) - X_2(t)] dt \end{aligned}$$

Unambiguously, $X_2 \rightarrow 0$ as $t \rightarrow \infty$, but $X_1 \rightarrow \infty$ as $t \rightarrow \infty$ unless

$$(38) \quad \begin{aligned} &\frac{1-\theta}{\theta}[AB(\mu) - \rho - \delta] < \rho \\ \Leftrightarrow &\frac{1-\theta}{\theta}[AB(\mu) - \rho - \delta] + \delta < \delta + \rho \end{aligned}$$

We therefore assume that this last inequality is satisfied, which gives us the inequality chain

$$(39) \quad \frac{1-\theta}{\theta} [AB(\mu) - \rho - \delta] + \delta < \delta + \rho < AB(\mu)$$

The first part guarantees bounded lifetime utility, and the second part guarantees positive growth.

Using the solution (36) for C_t , we have

$$(40) \quad \begin{aligned} dK^*/dt &= [AB(\mu) - \delta - \rho]K_t^* - C_0 e^{\frac{1}{\theta}[AB(\mu) - \delta - \rho]t} \\ \Rightarrow K_t^* &= \omega e^{[AB(\mu) - \delta]t} + \frac{C_0}{\xi} e^{\frac{1}{\theta}[AB(\mu) - \delta]t} \end{aligned}$$

where

$$(41) \quad \begin{aligned} \omega &= \text{constant of integration} \\ \xi &= \frac{[AB(\mu) - \delta](\theta - 1)}{\theta} + \frac{\rho}{\theta} > 0 \end{aligned}$$

and the last inequality is guaranteed by (38). From the equation for $d\psi/dt$ we have

$$(42) \quad d\psi/dt = \psi_0 e^{-[AB(\mu) - \delta - \rho]t}$$

Transversality requires

$$\begin{aligned} \lim_{t \rightarrow \infty} [K_t^* e^{-[AB(\mu) - \delta - \rho]t}] &= 0 \\ \Rightarrow \lim_{t \rightarrow \infty} [\omega + \frac{C_0}{\xi} e^{-\xi t}] &= 0 \\ \Rightarrow \omega &= 0 \quad \text{because } \xi > 0 \\ \Rightarrow K_t^* &= \frac{C_0}{\xi} e^{\frac{1}{\theta}[AB(\mu) - \delta - \rho]t} \end{aligned}$$

We thus have that consumption is proportional to the capital stock:

$$(43) \quad C_t = \xi K_t^*$$

from which we conclude that the growth rates of C and K^* are equal. Also, the growth rate of aggregate output equals the growth rate of aggregate capital because of the linearity of the production function (17) in K^* . In summary, all growth rates are equal:

$$(44) \quad \gamma_C = \gamma_{K^*} = \gamma_{Y^*} = \frac{1}{\theta} [AB(\mu) - \delta - \rho] \equiv \gamma$$

Two important characteristics of the common growth rate γ are that it is constant over time and it is a function of μ .

Constancy of γ over time means that there are no transition dynamics. Any shock moves the economy instantly to its new balanced growth path (the dynamic equivalent of the steady state). For example, a permanent increase in A raises ξ and thus C_0 and also raises γ . However, the economy jumps to its new balanced growth path with no transition, unlike the behavior one sees in a Cass-Ramsey model of aggregate growth.

The dependency of γ on μ means that the extent of monitoring affects the growth rate of the economy, not just the level of output. We showed earlier that $B'(\mu) = G(\mu) - g'(\mu)$. It is clear from Figure 3 that $B'(\mu) \geq 0$ as $\mu \leq \mu^*$. Therefore, the economy's growth rate γ increases in μ if $\mu < \mu^*$ and decreases in μ if $\mu > \mu^*$. It is maximized at $\mu = \mu^*$.

The response of consumption to changes in μ is somewhat surprising. From (41) and (43), we obtain the response of C_0 to μ :

$$(45) \quad \begin{aligned} dC_0/d\mu &= \frac{\theta-1}{\theta} AB'(\mu) K_0^* \\ &\begin{cases} \geq 0 & \text{as } (\theta-1)B'(\mu) \geq 0 \\ < 0 & \text{as } (\theta-1)B'(\mu) < 0 \end{cases} \end{aligned}$$

But

$$(46) \quad \begin{aligned} B'(\mu) &\begin{cases} \geq 0 & \text{as } \mu \leq \mu^* \\ < 0 & \text{as } \mu > \mu^* \end{cases} \\ \theta-1 &\begin{cases} \geq 0 & \text{as } \theta \geq 1 \\ < 0 & \text{as } \theta < 1 \end{cases} \end{aligned}$$

So C_0 responds to a change in μ with an instantaneous jump, but the direction of the response depends on the magnitudes of both μ and θ . If $\theta < 1$, then a movement of μ toward μ^* always

reduces initial consumption C_0 , irrespective of whether μ is moving up from a value initially below μ^* or is moving down from a value initially above μ^* . Conversely, if $\theta > 1$, a movement in μ toward μ^* always raises C_0 . In all cases, however, a movement of μ toward μ^* raises the growth rates of consumption, capital, and output.

3. *Capital Regulation.* In our analysis, there is no role for financial regulation. The social planner chooses the socially optimal amount of intermediation directly by allocating capital between production and monitoring to maximize output. However, if one were to extend the analysis to allow a role for regulation, one would need to proceed within the kind of general equilibrium framework used above, analyzing the allocation of physical capital between the production and intermediary sectors. Such an approach suggests a new orientation for thinking about the regulation of financial institutions' capital ratios. The discussion of capital adequacy requirements generally is couched in terms of which financial assets belong in the required capital ratios, how to adjust for their risk characteristics, and so on. Our type of analysis addresses none of those issues. In the central planning version we have presented here, all concern centers on the allocation of physical capital, and the financial sector's capital structure does not even exist. Nonetheless, we suggest that our approach is the right place to start thinking about regulating capital ratios. In general equilibrium, financial asset ratios and regulation affect the allocation of physical capital. Regulation changes not only the allocation of financial assets but also the allocation of the corresponding physical capital. Ultimately, it is the allocation of physical assets that is important to economic activity, so the first concern in evaluating financial regulation should be how it affects that allocation. Other issues are of secondary importance.

V. Conclusion

In this paper, we have addressed the theoretical question of what is the optimal size of the financial sector. This question recently has become especially important for Europe, in light of the continuing financial integration taking place there, as most recently evidenced by the advent of the Euro, a single currency for much of the continent. Proceeding in a general equilibrium framework that is an extension of Holmstrom and Tirole's partial equilibrium model, we have shown that there is indeed a unique optimal size for the financial sector. We derive the conditions necessary to determine the optimal size of the financial sector relative to the size of the economy as a whole. Creating and maintaining this sector requires diversion of some physical capital from production of output to monitoring that production, but the efficiency gain in output production brought about by monitoring warrants the diversion.

Some implications of our model are quite different from Holmstrom and Tirole's, even though our model is based on theirs. In particular, we find that the optimal size of the financial sector is independent of the state of the economy and does not vary over the business cycle. Also, we are able to address issues beyond the scope of their model, such as the effect of intermediation on the level and growth rate of aggregate output and on the behavior of consumption.

We suspect our conclusions on the acyclicity of the financial sector's optimal size arise from the AK type of production function that we have used and would not hold in a growth model with transition dynamics, such as the Lucas-Uzawa two-sector model or a one-sector growth model with a CES or Jones-Manuelli production function.⁶ Extending our work to such models would be useful. Whatever the outcome of such extensions may be, the general

⁶See chapters 4 and 5 of Barro and Sala-i-Martin, 1995, for a discussion of those models.

equilibrium framework we have used here is necessary if one is to address the kinds of questions that must be asked in any attempt to regulate the financial sector. Even our simple model shows how incorrect choice of the financial sector's capital ratio, $\mu = K_m^*/K^*$, has adverse consequences for aggregate output, investment, consumption, growth rates, and social welfare.

References

- Allen, Franklin and Douglas Gale. "Optimal Security Design," *Review of Financial Studies*, 1988, v1(3), 229-263.
- Allen, Franklin and Anthony M. Santomero, "The Theory of Financial Intermediation," *Journal of Banking and Finance*, 1997, v21, 1461-1485.
- Barro, R.J., and X. Sala-i-Martin (1995). *Economic Growth*. McGraw-Hill, New York.
- Bhattacharya, Sudipto and Anjan V. Thakor. "Contemporary Banking Theory," *Journal of Financial Intermediation*, 1993, v3(1), 2-50.
- Bernanke, Ben S. "Bankruptcy, Liquidity, And Recession," *American Economic Review*, 1981, v71(2), 155-159.
- Bernanke, Ben and Mark Gertler. "Financial Fragility And Economic Performance," *Quarterly Journal of Economics*, 1990, v105(1), 87-114.
- Bernanke, Ben and Mark Gertler. "Agency Costs, Net Worth, And Business Fluctuations," *American Economic Review*, 1989, v79(1), 14-31.
- Brunner, K. and A. H. Meltzer. "The Place of Financial Intermediaries In The Transmission Of Monetary Policy," *American Economic Review*, 1963, v53(2), 372-382.
- Diamond, Douglas W. "Financial Intermediation And Delegated Monitoring," *Review of Economic Studies*, 1984, v51(166), 393-414.
- Fama, Eugene F. "What's Different About Banks?," *Journal of Monetary Economics*, 1985, v15(1), 29-40.
- Holmstrom, B., and J. Tirole (1997). "Financial Intermediation, Loanable Funds, and the Real Sector," *Quarterly Journal of Economics* 112, August, pp. 663-691.
- Leland, Hayne E. and David H. Pyle. "Informational Asymmetries, Financial Structure, And Financial Intermediation," *Journal of Finance*, 1977, v32(2), 371-387.
- Patankin, Don. Money Interest and Prices, Row, Peterson (1956).
- Tobin, James. "A General Equilibrium Approach To Monetary Theory," *Journal of Money, Credit and Banking*, 1969, v1(1), 15-29.

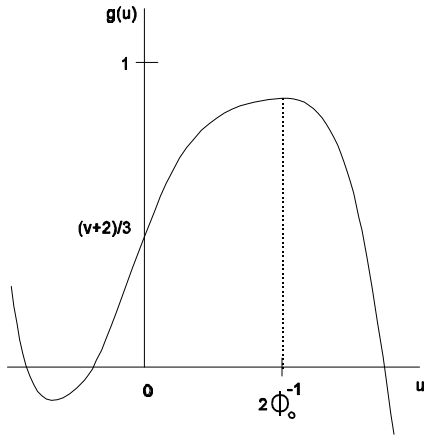


Figure 1: The g function

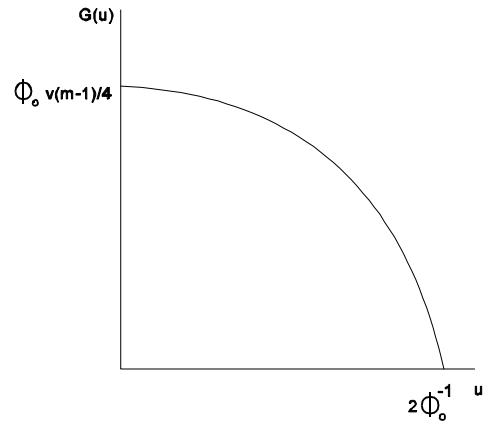


Figure 2: The G function

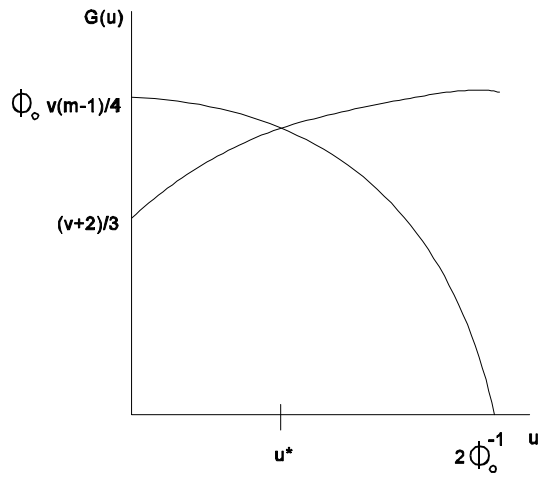


Figure 3: Optimal u