The Equilibrium Size of the Financial Sector

Thomas Philippon*
New York University
August 2007

Abstract
Over the past 60 years, the U.S. financial sector has grown from 3% to 9% of private GDP. I present a model of the equilibrium size of this industry and I study the factors that might explain why it has increased so much. Productivity gains in the financial sector and an increase in the corporate demand for financial services seem able to account for three quarters of the increase. According to the model, the demand shift comes from a decrease in the correlation between investment opportunities and current income across U.S. firms. An increase in productivity in the financial sector also seems to play an important role. Without productivity gains, the fraction of credit constrained firms would be almost 50% larger than it is today.

*Stern School of Business, NBER and CEPR.
Figure 1 shows the evolution of the share of the U.S. private valued added accounted for by the Finance and Insurance industry over the post-war period. This share has grown from 3% to more than 9%. Understanding this evolution is the main motivation of this paper. To do so, I present a simple general equilibrium model of a production economy and I study the factors that pin down the supply of and demand for financial services.

The model economy has overlapping generations of agents and two sectors: industrial and financial. In the first period of their lives, agents decide in which sector they wish to work. The industrial sector produces a good that can be consumed or invested. All agents in the industrial sector engage in production, albeit with different levels of productivity. Some of them also receive an investment opportunity. Borrowing is limited by moral hazard, and some valuable investments may not be financed. Agents in the financial sector have access to a monitoring technology that reduces moral hazard. They sell their services to agents who would not be able to invest otherwise.

A theory of the equilibrium size of the financial sector can shed light on a number of relevant issues. First, what determines the equilibrium demand for financial services? In particular, what is the role of the joint distribution of cash flows and growth options? Second, on the supply side, what are the consequences of productivity gains in the financial sector? Do they lead to a larger or a smaller finance industry? How do they affect the level of credit rationing? Do they necessarily lead to more investment? How can we actually measure these productivity gains? Third, what are the links between intermediated financing and direct financing? Should we expect them to grow simultaneously, or should we expect a substitution from one to the other? It is common in the empirical literature to measure financial development using private credit over GDP. Is this a good measure of financial development? How does it relate to the size of the financial sector measured with value added? Does an increase in the size of the financial sector signal that more firms are constrained, or fewer? When productivity improves, does the financial sector grow until all credit constraints disappear? To the best of my knowledge, no existing model addresses these issues in a simple unified framework.

The paper is related to the rich literature on financial intermediation, but my concern is with the macroeconomic outcome more that with the microeconomic ones. I therefore abstract from the issues of delegated monitoring emphasized in Diamond (1984), from the
supply of bank capital studied by Holmström and Tirole (1997), and from the formation of optimal coalitions as in Boyd and Prescott (1986). In the model, the cost of financial intermediation is an opportunity cost, because an agent cannot be a banker, an engineer or a worker at the same time. The paper is also related to the literature that studies the link between financial intermediation and growth: Bencivenga and Smith (1991), Greenwood and Jovanovic (1990), Levine (1991) among others. More recently, Greenwood, Sanchez, and Wang (2007) study the quantitative impact of efficiency gains in monitoring on economic performance in a cross-section of countries. Finally, the paper is related to the literature on economic growth and structural change, with early contributions by Stigler (1956), Kuznets (1957), and Baumol (1967). The rise in the finance industry depicted here is similar to the rise in some other skill-intensive services, as shown recently by Buera and Kaboski (2006). The literature on structural change focuses on non-homothetic preferences and the trade-off between home production and market production. My approach instead relies on a standard model of financing constraints, augmented with endogenous monitoring and career choices.

Section 1 presents the model. Section 2 characterizes the equilibrium. Section 3 studies the determinants the equilibrium level of financial services provided in the economy. Section 4 presents a calibration and proposes an interpretation for the growth of U.S. financial sector since the 1950s. Section 5 concludes.

1 The model

1.1 Technology and preferences

Consider an economy with overlapping generations of agents who live for two periods. They work when they are young and discount the future at rate $\rho$. The size of each generation is normalized to 1. An agent $i \in [0, 1]$ born at time $t$ seeks to maximize her expected utility:

$$U^i_t = E_t \left[ c^i_{1,t} + \frac{c^i_{2,t+1}}{1 + \rho} \right].$$

(1)

The economy has two sectors, industrial and financial. The industrial sector produces a good that can be consumed or invested. The financial sector produces monitoring services that are used by entrepreneurs of the industrial sector. I now describe in more details the supply and demand of each good, as well as the investment process.
Career Choice

In their first period, agents choose a career. Let \( n_t \) be the number of agents who choose the industrial sector. The remaining \( 1 - n_t \) enter the financial sector. I start by describing an agent’s career within the industrial sector. After she enters the industrial sector, an agent receives two endowment shocks: \( \tilde{\theta} \in \{0, \theta\} \) and \( \alpha \in (0, \infty) \). Both shocks are publicly observable. The first shock measures the investment opportunity of the agent. Investment requires \( x \) units of consumption good at time \( t \) and delivers \( \tilde{\theta}x \) units of capital at time \( t + 1 \). Each unit of capital is a Lucas tree that depreciates at rate \( \delta \) and yields 1 unit of consumption good each period. Let \( \pi \) be the fraction of agents in the industrial sector who receive an investment opportunity:

\[
\pi \equiv \Pr(\tilde{\theta} = \theta).
\]

The second shock determines the productivity of the agent in the first period of her life. A agent who receives a shock \( \alpha \) produces \( \alpha x \) units of output (scaling by \( x \) is a normalization that simplifies the notations later on). Let \( \bar{\alpha} \) be the unconditional mean of \( \alpha \):

\[
\bar{\alpha} \equiv E[\alpha].
\]

The parameter \( \bar{\alpha} > 0 \) measures the average productivity of the industrial sector, relative to its physical capital requirements. The shocks \( \alpha \) and \( \tilde{\theta} \) are correlated. Let \( F(.) \) be the cumulative distribution of \( \alpha \) conditional on \( \tilde{\theta} = \theta \), and let \( f(.) \) be the density function. The distribution of \( \alpha \) conditional on \( \tilde{\theta} = 0 \) plays no role in the analysis and needs not be specified explicitly.

Production and capital accumulation

The consumption good is produced by agents in the industrial sector, with average productivity \( \bar{\alpha} \), and by the Lucas trees, with a yield of one unit of consumption good per unit of capital. Let \( k_t \) be the stock of capital at the beginning of period \( t \). The total amount of goods produced at time \( t \) is:

\[
y_t = \bar{\alpha}x n_t + k_t.
\]

Capital accumulates over time according to:

\[
k_{t+1} = (1 - \delta) k_t + \theta x e_t,
\]
where \( e_t \) is the number of agents who invest. Since only a fraction \( \pi \) of agents in the industrial sector have a positive innovation opportunity, \( e_t \) cannot be more than \( \pi n_t \).

**Enforcement constraint and monitoring**

Entrepreneurs have access to a stealing technology. After investing one unit at time \( t \), an entrepreneur can always default at time \( t + 1 \) and steal \( zx \) from her project, while the remaining \( \theta x - zx \) is destroyed. The parameter \( z \in (0, \theta) \) captures the severity of the moral hazard problem. Moral hazard may prevent the entrepreneur from obtaining the necessary funds for the initial investment. To alleviate the financing constraint, the entrepreneur can hire monitoring services. If an entrepreneur hires \( m \) units of monitoring services, the amount she can steal is reduced to \( (z - m)x \). The monitoring services are supplied by the financial sector. Each agent in the financial sector can supply up to \( \mu \) units of services. The parameter \( \mu \) measures the productivity of the financial services industry. Since the number of agents born each period is normalized to one, the total amount of monitoring available in the economy at time \( t \) is:

\[
\mu (1 - n_t).
\]

### 1.2 Discussion of the model

The model has several special features, which were chosen to make the analysis transparent and tractable. The first main assumption concerns the role of the financial sector. Levine (2005) defines five broad functions of the financial sector: (i) to produce information and allocate capital ex-ante, (ii) to monitor investments and exert corporate governance, (iii) to facilitate trading and diversification, (iv) to mobilize and pool savings and (v) to ease the exchange of goods and services. In this paper, I focus on the monitoring of corporate investment and the allocation of capital. The demand for financial services is entirely due to moral hazard, and I abstract from transaction costs and trading frictions.

Broadly speaking, there are two approaches to modelling financial services. One can assume the presence of transaction costs and study the organization of the industry. In this approach, financial institutions (FIs) are to financial products what retailers are to goods and services. However, as Freixas and Rochet (1997) argue, “the progress experienced recently in telecommunications and computers implies that FIs would be bound to
disappear if another, more fundamental, form of transaction costs were not present”. A second approach to modelling FIs focuses instead on information asymmetries and moral hazard. Modeling financial intermediaries as monitors has a long tradition in economics and finance, and much work has been done on the issue of who monitors the monitors (Diamond (1984), Holmström and Tirole (1997)). I abstract from this issue by not introducing any asymmetric information or moral hazard between savers and financial intermediaries. This implies that firm boundaries within the financial sector are irrelevant. All that matters is the productivity of the sector, measured by the parameter $\mu$.

Regarding preferences, the main assumption is that agents are risk neutral. As explained in the introduction, this paper looks at the role of the financial sector in providing services to the corporate sector. The paper does not take into account the services provided to households, in terms of liquidity and personal insurance. Risk neutrality also implies that the real rate of interest is constant as long as consumption is interior for at least some agents.

On the production side of the model, the main assumption is that $k$ and $n$ since are perfect substitute in equation (2). This implies a fixed relative price, and it simplifies the analysis by reducing the dimension of the system. Notice that one should be careful in interpreting $n$ as labor and $k$ as capital, however. In the model, $\alpha$ pins down the current cash flows of agents, who can also be viewed as firms. Conversely, the entrepreneurs receive the net present value of their investment option, which adds to the future stock of $k$, but in practice it would partly show up as labor income.

Finally, the model assumes a closed economy. This means that the entire demand for financial services comes from domestic firms. I return to this issue in interpreting the results of the calibration in section 4.

2 Equilibrium

Market clearing

Equilibrium in the goods market requires that consumption plus investment equals total production. Let $c_{1,t}$ be the total consumption of the young agents, and let $c_{2,t}$ be the total
consumption of the old agents. The market clearing condition is:

\[ c_{1,t} + c_{2,t} + xe_t = k_t + \bar{\alpha}xn_t. \]  (3)

In their second period, old agents receive dividends equal to \( k_t \). They also own the stock of capital that remains after depreciation. Let \( q_t \) be the ex-dividend price of one unit of capital. Without bequest motives, the second period consumption of the old generation equals its total income:

\[ c_{2,t} = k_t + (1 - \delta) q_t k_t. \]  (4)

Note that \( c_{1,t} \) and \( c_{2,t} \) are aggregate quantities and that agents within a generation typically have different levels of consumption. Combining (3) and (4), we obtain the saving equation:

\[ \bar{\alpha}xn_t - c_{1,t} = (1 - \delta) q_t k_t + xe_t. \]  (5)

The left-hand-side of equation (5) measures the savings of the young generation. The right-hand-side of the equation measures the investments of the young generation, which consist of buying the existing machines from the old generation, and financing new projects.

**Asset prices**

Let \( r_t \) be the interest rate between period \( t \) and \( t + 1 \). The ex-dividend price of one unit of installed capital satisfies the dynamic equation:

\[ q_t = \frac{1 + (1 - \delta) q_{t+1}}{1 + r_t}. \]  (6)

Let \( xv_t \) be the net present value of a new investment. The net present value per unit satisfies

\[ v_t = \theta \frac{1 + (1 - \delta) q_{t+1}}{1 + r_t} - 1. \]  (7)

**Steady State**

I focus on the long run equilibrium of the economy, which I associate with the steady state of the model. When the interest rate is constant, the value of a project is:

\[ v = \frac{\theta}{r + \delta} - 1. \]  (8)

In steady state, the level of capital is \( k = \theta ex/\delta \) and the ex-dividend price of one capital becomes \( q = 1/ (r + \delta) \). The saving equation (5) becomes

\[ c_1/x = \bar{\alpha}xn - g (r) e, \]  (9)
where the function \( g(r) \) is defined by:
\[
g(r) \equiv 1 + \frac{\theta}{r + \delta} \frac{1 - \delta}{\delta}.
\]

In equilibrium, the interest rate and the investment demand must adjust so that \( c_1 \) is always positive.

I now describe the steady states of three economies. First, I characterize the equilibrium of an economy without enforcement constraints. Second, I describe the equilibrium of an economy with moral hazard and without intermediation. The discussion of these two benchmarks is brief and is mainly used to provide the foundations for the analysis of the third economy, where moral hazard and active monitoring are both present.

### 2.1 No moral hazard

In this section, I consider the case where \( z = 0 \). Projects are funded if and only if they are profitable, which is determined by equation (8). There are no bankers, and \( n = 1 \). The investment demand (in units of entrepreneurship) is therefore
\[
ed^d(r) = \begin{cases} 
\pi & \text{if } r + \delta < \theta \\
[0, \pi] & \text{if } r + \delta = \theta \\
0 & \text{if } r + \delta > \theta
\end{cases}
\]

The saving curve, derived from equation (9) and the constraint that \( c_1 \) be positive, is
\[
es^s(r) = \begin{cases} 
0 & \text{if } r < \rho \\
[0, \bar{\alpha}/g(\rho)] & \text{if } r = \rho \\
\bar{\alpha}/g(r) & \text{if } r > \rho
\end{cases}
\]

The interest rate is pinned down by the market clearing condition: \( e^s(r) = e^d(r) \). The following proposition characterizes the first best equilibrium, which is also depicted on Figure 2:

**Proposition 1** Without moral hazard, capital investments are undertaken if and only if \( \theta > \rho + \delta \). The equilibrium interest rate depends on average productivity and capital requirements. If \( \pi \leq \delta \bar{\alpha} \), then \( r \in [\rho, \theta - \delta] \) and all investment projects are financed. If \( \pi > \delta \bar{\alpha} \), then \( r = \theta - \delta \), entrepreneurs are indifferent between investing and not investing, and \( e < \pi \).

One important feature of this economy is that the equilibrium is independent of the conditional distribution of income \( F(.) \). Only the unconditional mean \( \bar{\alpha} \) matters. From now on,
I assume that the no-moral hazard economy has a strictly positive investment rate with an interior solution for consumption. This holds under the following parameter restriction:

**Assumption 1:** \( \theta > \rho + \delta \) and \( \bar{\alpha} > \pi g (\rho) \)

### 2.2 Moral hazard without monitoring

I now consider the case where \( z > 0 \), but I still assume that there is no active financial sector (one can think of this case as \( \mu = 0 \)). Borrowing is limited by the fact that the entrepreneur can default. Consider an entrepreneur who has borrowed an amount \( b \) in her first period. Since each unit of capital (before depreciation) yields one unit of dividend each period, the cum-dividend value per unit is

\[
\frac{(1 + r)}{(r + \delta)}.
\]

If the entrepreneur defaults, she gets \( zx \frac{(1 + r)}{(r + \delta)} \) in the second period of her life. If she does not default, she gets \( (1 + r) \theta x \frac{1}{(r + \delta)} - b (1 + r) \). The maximum amount of borrowing allowed in the first period is therefore \( b^{\text{max}} = (\theta - z) x \frac{1}{(r + \delta)} \). An entrepreneur with current income \( \alpha x \) can finance her investment if and only if \( \alpha x + b^{\text{max}} > x \), which leads to the financing constraint:

\[
\alpha > 1 + \frac{z - \theta}{r + \delta}.
\]

An entrepreneur can invest if and only if \( \alpha > \alpha_h \), where the threshold for financing without monitoring is defined by:

\[
\alpha_h \equiv \frac{z}{r + \delta} - v.
\]

The threshold is equal to the gap between the part of the project that cannot be pledge to outside investors and the excess return on the project. Note that this threshold depends on the interest rate.

Let \( e_c (r) \) be the effective investment demand curve under moral hazard. When \( r + \delta > \theta \), it collapses to zero, just like in the case of no moral hazard examined in the previous section. When \( r + \delta \leq \theta \), the constrained investment demand is given by:

\[
e_c (r) = \pi (1 - F (\alpha_h)).
\]

When \( r + \delta = \theta \), the effective demand curve is vertical and \( e_c \) can be anywhere between 0 and \( \pi (1 - F (\frac{\theta}{2})) \). The saving equation is still given by (11), as in the case of no moral hazard.

The equilibrium is depicted on Figure 3 and characterized by the following proposition:
Proposition 2 Under assumption 1, in the equilibrium with moral hazard and no monitoring, the interest rate is \( \rho \) and the number of project financed is \( e^c(\rho) \).

Proof. Under assumption 1, equation (9) shows that \( c_1 > 0 \). Therefore \( r = \rho \). Investment is pinned down by the constrained demand schedule \( e^c \) evaluated at \( r = \rho \). □

2.3 Active monitoring

I now turn to the case where firms have the option to hire monitoring services in order to relax their credit constraints. Suppose that an entrepreneur with current income \( \alpha \) hires \( m \) units of monitoring. The financing constraint (12) becomes: \( \alpha + vx \geq (z - m)x / (r + \delta) \).

The amount of monitoring required for this entrepreneur is therefore

\[
m(\alpha) = (r + \delta)(\alpha_h - \alpha). \tag{15}
\]

Let \( \phi \) be the price of one unit of monitoring. For intermediation to be valuable to entrepreneurs, the price of financial services must be such that \( vx - m\phi \geq 0 \). It is profitable for an entrepreneur to hire monitoring services if \( \alpha \) is more than \( \alpha_l \), defined as the solution to:

\[
m(\alpha_l) = \frac{vx}{\phi}. \tag{16}
\]

Let us now turn to the supply of financial services. In their first period, agents choose freely which sector they would like to work. Therefore, in any equilibrium where at least some agents become financial intermediaries, the following indifference condition must hold:

\[
\mu \phi = \bar{\alpha}x + \pi (1 - F(\alpha_h))vx + \pi \int_{\alpha_l}^{\alpha_h} (vx - \phi m(l)) dF(\alpha). \tag{17}
\]

The left-hand-side of equation (17) is the value of entering the financial sector. The right-hand-side is the value of entering the non-financial sector, which contains three terms. The first term is the expected income received when working in the industrial sector. The second term is the expected value of becoming an unconstrained entrepreneur. The last term is the expected value of becoming a constrained entrepreneur who hires financial services.

Combining (16) and (17), I obtain:

\[
\frac{\mu}{\pi} = \left( \frac{\bar{\alpha}}{\pi v} + 1 - F(\alpha_l) \right) m(\alpha_l) - \int_{\alpha_l}^{\alpha_h} m(\alpha) dF(\alpha) \tag{18}
\]
Finally, equilibrium in the monitoring market requires that

$$\frac{1 - n \mu}{n \pi} = \int_{\alpha_l}^{\alpha_h} m(\alpha) dF(\alpha).$$

(19)

The difference with the previous case of moral hazard without monitoring is that the saving curve in Figure 3 also shifts as people move in and out of the industrial sector. Figure 4 describes the monitoring equilibrium.

**Proposition 3** Under assumption 1, the equilibrium with active monitoring is characterized by \((\alpha_l, n)\) that solve equations (18) and (19) for \(r = \rho\). Credit rationing persists in equilibrium as long as \(\alpha_l > 0\). The size of the financial sector is strictly positive for all values of \(\mu \in (0, \infty)\), provided that the density is strictly positive at the self-financing threshold \(\alpha_h\).

**Proof.** Under assumption 1, equation (9) shows that \(c_1 > 0\). Therefore \(r = \rho\). The RHS of equation (18) goes to zero as \(\alpha_l \to \alpha_h\) so for any value of \(\mu > 0\), it is possible to find \(\alpha_l < \alpha_h\) that solves equation (18). If the density \(f(\alpha_h) > 0\) and if \(f\) is continuous, then the RHS of (19) is strictly positive and \(n\) is strictly less than one. QED.

Even when \(\mu\) is very small, some agents always choose to become financial intermediaries. Of course, this means that \(\phi\) must become arbitrary large. But close to the cutoff \(\alpha_h\), the amount of monitoring required to obtain financing is arbitrary small, and some entrepreneurs are always willing to buy this small amount, even at a very high price. Therefore, the banking sector is active for any positive value of \(\mu\).

### 3 Comparative statics

I organize the comparative statics in the following way. I first study the parameters specific to moral hazard and financial intermediation, i.e. the parameters that affect the equilibrium of the economy with intermediation, but do not appear in the equilibrium of the economy without moral hazard. I then study the parameters that determine the macroeconomic investment opportunity set.

Three parameters affect the supply and demand of financial services in the model economy with moral hazard and intermediation, but would be irrelevant in an economy without
moral hazard. On the supply side, $\mu$ measures the productivity of the financial sector. On the demand side, $z$ measures the degree of moral hazard, while $F(.)$ measures the extent to which firms with growth options also have high current income.

Let us consider the supply side first. How does the size of the financial sector depend on $\mu$? The following two propositions characterize the effects of $\mu$. Consider first the extreme cases where the productivity of the financial sector is either very high or very low. Looking at equation (19), it seems difficult to predict the limiting behavior of the size of the financial sector, since the quantity of monitoring on the right hand side goes to zero when $\mu$ goes to zero. The behavior of $n$ is therefore unclear. It turns out, however, that one can prove a general result, that holds for all distribution functions:

**Proposition 4** The size of the financial sector goes to zero when its productivity becomes either very small or very large.

**Proof.** See appendix. ■

On the one hand, an increase in $\mu$ implies that the same amount of monitoring can be performed by fewer bankers. On the other hand, the drop in the price of monitoring services leads to a surge in demand. These two forces determine the effects of changes in the productivity of monitoring services. When $n$ is close to one, the supply effect is negligible and the demand effect dominates. Therefore, starting from a value of $\mu$ close to zero, which implies a value of $n$ close to one, as we have shown in the previous proposition, we see that an increase in $\mu$ leads to a decrease in $n$. For a very large value of $\mu$, the density $f(\alpha_l)$ must eventually be close to zero. Therefore, $n$ must increase in response to an increase in $\mu$. For intermediate values of $\mu$, the comparative statics obviously depend on the shape of the density function $f$. The demand effect dominates if the density at $\alpha_l$ is high enough, and an increase in $\mu$ then leads to an increase in the size of the financial sector. Finally, note that the increase in $n$ might occur even though there is still a positive density at the cutoff $\alpha_l$. Therefore the financial sector might shrink despite the fact that some firms remain rationed. Of course, one should keep in mind that the supply of monitoring services rises in any case, since $\alpha_l$ always decreases with $\mu$. 

12
Proposition 5  Productivity gains in the financial sector always reduce credit rationing. Productivity gains lead to larger financial sector when the productivity of the financial sector is relatively low. When the productivity of the financial sector is already high, further productivity gains decrease the size of the sector. In all cases, it is possible for the size of the financial sector to decrease in response to productivity gains even though some firms remain rationed.

Proof. See appendix.

Consider now a change in $z$. The change in $z$ affects the self-financing cutoff $\alpha_h$. This changes the monitoring function as well as the boundaries of integration. As expected, when moral hazard worsens, credit rationing increases. The impact on the size of the financial sector is ambiguous, however. On the one hand, it takes more resources to monitor a given set of firms. On the other hand, the pool of firms that are actually monitored shrinks.

The second parameter that determines the demand for financial services is the distribution function $F$. Remember that $F$ is the distribution of $\alpha$ conditional on $\tilde{\theta} = \theta$. Thus, a change in $F$ while keeping the unconditional mean $\bar{\alpha}$ constant, is like changing the cross-sectional correlation between investment opportunities and current cash flows, while keeping aggregate productivity constant. The comparative statics for $F$ are unambiguous, unlike those for $z$. A decrease in $f$, that is, a shift of $F$ to the right must always increase $\alpha_l$. This is because the shadow value of an extra unit of income, conditional on having an investment opportunity, is more than one. Therefore, a right shift in $F$, even holding $\bar{\alpha}$ constant, makes it more attractive to work in the industrial sector. To keep agents indifferent between the two sectors, the price of financial services must increase, and $\alpha_l$ must increase. This reduces the demand for financial services in two ways. It shrinks the range of values of $\alpha$ for which monitoring is purchased, and it decreases the number of firms who need monitoring for a given monitoring region.

Proposition 6  Worsening of moral hazard increases credit rationing but has an ambiguous effect on the size of the financial sector. A decrease in the correlation between current income and investment options decreases $\alpha_h$ and increases the size of the financial sector.

Proof. See appendix.
I now turn to the parameters that affect the investment opportunity set: $\pi$ and $\theta$. A shift in $\pi$ is like a broadening of the investment set: more firms can invest, but the average productivity does not change. The increase in $\pi$ increases the attractiveness of the industrial sector, therefore the price of financial services must rise to satisfy the indifference condition. This means that $\alpha_l$ must increase. Thus, rationing increases. The impact on $n$ is ambiguous, since more firms would like to invest, but a smaller fraction can actually do so. A shift in $\theta$ is like a deepening of the investment set: each project becomes more valuable, but the number of projects stays constant. Clearly, less monitoring is required per firm and $\alpha_h$ decreases. The impact on $n$ is ambiguous, since more firms are monitored, but conditional on $\alpha$ monitoring decreases.

**Proposition 7**  
An increase in $\theta$, or a decrease in $\pi$, decreases rationing and has ambiguous effects on the size of the financial sector

**Proof.** See appendix. ■

4  
A quantitative investigation

4.1 Calibration of the parameters

The data used in the calibration is presented in Table 1. It covers the period 1956-2005, split into five sub-samples. I start by choosing standard values for the discount rate and for the depreciation rate:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Empirical Value</th>
<th>Model Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of 1 Period</td>
<td>15 years</td>
<td>$\rho = 1.02^{15} - 1$</td>
</tr>
<tr>
<td>Annual real rate</td>
<td>2%</td>
<td>$\delta = 1 - 0.92^{15}$</td>
</tr>
<tr>
<td>Annual depreciation rate</td>
<td>8%</td>
<td></td>
</tr>
</tbody>
</table>

Next, I use the fact that the book value of a realized project is $x$ while its market value is $\theta x / (\rho + \delta)$. Using a ratio of market value to book value of 2,\(^1\) I obtain $\theta = 2 (\rho + \delta)$. It seems *a priori* difficult to calibrate the remaining parameters of the model, since they are not observable. Notice, however, that the equilibrium equations (18) and (19) depend

\(^{1}\)This is a typical value for the US over the post-war period. Fama and French (2001) report an asset weighted average of 1.4 over the period 1963-1998. Hennessy, Levy, and Whited (2007) calibrate their model with an average value of 2.5 over 1968-2003, and they report values above 3 for firms that either issue equity or are likely to be credit constrained.
only on the ratios $\mu/\pi$ and $\bar{\alpha}/\pi$. To further reduce the number of degrees of freedom in the calibration, I consider families of functions $f(.)$ that depend on only one parameter. I perform the calibration using either a uniform distribution or a downward sloping triangular distribution, both on the interval $[0, \alpha^{\text{sup}}]$. The unknowns are therefore: $z$, $\mu/\pi$, $\bar{\alpha}/\pi$ and $\alpha^{\text{sup}}$. I now explain how I use four observed quantities to pin down these parameters.

**Capital Expenditures**

In the model economy, aggregate investment expenditures are equal to $(1 - F(\alpha_l)) \pi n x$, and the equilibrium stock of Lucas trees is:

$$k = (1 - F(\alpha_l)) \pi n x \theta / \delta.$$  

The gross domestic product is $\bar{\alpha} n x + k$. Therefore the investment share of GDP depends only on $\alpha_l$ and on the ratio $\bar{\alpha}/\pi$:

$$\frac{1 - F(\alpha_l)}{\bar{\alpha}/\pi + (1 - F(\alpha_l)) \theta / \delta}$$

I calibrate this ratio to 12.44% using private non-residential fixed investment divided by private value added from the National Income and Product Accounts.

**Investment share of low cash firms**

Recall that $x$ is capital required for a project, while $\alpha x$ is the current income of the agent. Thus, $\alpha$ captures the ratio of income to capital expenditures, for the firms that actually invest. To make progress here, I need to use some information about the distribution of firms and investments. In the model, for tractability, I have assumed a unique fixed scale $x$, but this is not a realistic assumption. I use a simple statistic to compare the model and the data: the share of total investment accounted for by firms whose income is less than one third of their capital expenditures. This statistic is useful because it does not involve taking a ratio of income over capital expenditures. It is robust to the discrepancy between the assumptions made in the model and the fact that firm sizes are very heterogenous in the real world. It also builds in the fact that large firms are more relevant than small firms. In the model, the investment share of low cash firms is $(F(1/3) - F(\alpha_l)) / (1 - F(\alpha_l))$. I use Compustat to estimate this share over time. I use all firms in the industrial Compustat files.
with non missing values for income before extraordinary items and capital expenditures, and I exclude financial firms and firms real estate. I compute:

\[ s_t = \frac{\sum_i capex_{it} \ast (income_{it} < capex_{it}/3)}{\sum_i capex_{it}} \]

The share of investment accounted for by firms with \( \alpha < 1/3 \) is shown on Table 1. It is has increased over time.\(^2\) The value of 1/3 is arbitrary. To be sure, the value must be substantially less than one because the relevant information is in the left of the distribution. To check the robustness of the calibration, I have performed the same exercise using 1/5 and 1/4 instead of 1/3, and obtained similar results.

**Corporate borrowing**

In the model, firms with \( \alpha < \alpha_t \) cannot invest, and firms with \( \alpha > 1 \) can finance their investment entirely with their current income, as depicted on Figure 4. Total corporate borrowing in each period is therefore equal to:

\[ \pi nx \int_{\alpha_t}^{1} (1 - \alpha) f(\alpha) d\alpha \]

I compute the ratio of outstanding credit market instruments over private value added from the Flow of Funds Accounts of the U.S. This ratio is shown in Table 1. It has increased over time. The relevant variable for the calibration is the amount of new borrowing in each period. I assume that the average maturity of credit market instruments is 10 years, and I calibrate the model to a borrowing ratio equal to 1/10 of the outstanding value from Table 1. For robustness, I have redone the analysis assuming average maturities from 8 to 15 years.

**Estimation**

Finally, I match \( 1 - n \) as the value added share of the financial sector, also from Table 1. I estimate the baseline parameters using the average values over the first two decades, 1956

---

\(^2\)One might worry about a change in the coverage of Compustat. However, after 1975, the ratio of the number of employees covered by Compustat to total non-farm payrolls is constant. It is true that these firms are younger, but this is exactly what one would expect if financial markets improve. See Philippon and Sannikov (2007).
to 1975. The estimated values, assuming that $f$ is uniform, are:

$$
\begin{array}{cccc}
\frac{z}{\theta} & \frac{\mu}{\pi} & \frac{\bar{\alpha}}{\pi} & \alpha^{\sup} \\
0.81 & 2.41 & 4.1 & 1.11
\end{array}
$$

The estimated values, assuming that $f$ is downward sloping triangular, are:

$$
\begin{array}{cccc}
\frac{z}{\theta} & \frac{\mu}{\pi} & \frac{\bar{\alpha}}{\pi} & \alpha^{\sup} \\
0.79 & 1.73 & 3.66 & 1.59
\end{array}
$$

4.2 Quantitative Properties

As explained above, the model is calibrated using values for the period 1956-1975. In this section, I investigate the quantitative properties of the model, in two ways. First, I compare the equilibrium for different values of the monitoring productivity parameter $\mu$. Second, I create predicted values for the period 1976-2005 and I compare the predictions with the actual values.

Impact of productivity gains

Figure 5 shows the impact of changing the productivity of the financial sector. All the parameters of the model are the ones calibrated for the period 1956-1975 assuming a uniform density $f$. The horizontal axis is normalized by the benchmark value for $\mu$, called $\mu_0$. The left panel shows the size of the financial sector, $1 - n$. The right panel shows the fraction of firms that are credit constrained. Consistent with proposition 5, Figure 5 shows that the size of the financial sector increases with $\mu$ when $\mu$ is small, and decreases with $\mu$ when $\mu$ is large. The right panel shows the fraction of firms that are credit constrained. The calibrated value is 13.24%. Productivity gains in the financial sector decrease credit rationing, as explained in proposition 5. While this is virtually undetectable, the size of the financial sector actually starts to decrease before the fraction of constrained firms reaches zero. This is qualitatively consistent with 5, but it appears quantitatively irrelevant. To a good approximation, productivity gains in finance should increase the size of the financial sector until all credit constraints are alleviated. Further gains are used to reallocate labor to the industrial sector while keeping enough financiers to make sure all positive NPV projects are financed.

Out-of-sample predictions
How well does the model predict the evolution of the economy from 1976 to 2005. I construct predicted values while keeping productivity $\tilde{\alpha}$, investment opportunities $(\pi, \theta)$ and moral hazard $z$ constant. I let $\alpha_t^{\text{sup}}$ and $\mu_t$ change over time and I pin down their values in the following way. First, I keep the investment share of GDP constant, which appears to be consistent with the evidence.\footnote{In the model, without changing any other technological parameter, this implies that the fraction of constrained firms also stays constant.} Second, I match the share of investment by low cash firms from Table 1. I then compare the predicted size of finance and corporate borrowings with the actual values from Table 1. Figure 6 shows the predictions of the model for the size of the financial sector. Figure 7 shows the predictions of the model for the ratio of credit market instruments over GDP. Up to the 1990s, the model seems able to predict most of the changes in the size of the financial sector. The model falls short in the more recent decade. This might be because of globalization or household finance, since the calibration forces all the demand for financial services to come from US firms.

Counter-factual experiments

Suppose there had been no shift in productivity or in the correlation between growth options and cash flows. What would have happened? Table 2 shows the results of counter-factual experiments using the calibrated model. Notice that the model is non linear, so the effects are not additive: the total change is not the sum of the two partial counter-factual changes. The fraction of constrained firms does not change over time because I have calibrated the model to keep a constant investment share of GDP. In terms of size, most of the increase appears to come from a shift in the correlation between investment opportunities and current income. Even without productivity gains, the model predicts an increase in the size of the industry by 3.3 percentage points of GDP. Productivity gains only add another 0.55 percentage points. However, without productivity gains, the fraction of rationed firms would have increased substantially, from 13.24% to 19.57%.

5 Conclusion

This paper is a first attempt at building a quantitative model of the size of the financial sector. The model sheds light on a number of issues, such as the consequences of productivity
growth in the finance industry or the determinants of the corporate demand for financial services. In the model, the equilibrium size of the financial sector depends on two main parameters: the productivity of the monitoring technology used by the financial sector, and the correlation across industrial firms between growth options and current income. Because the efficiency of the financial sector affects the realized correlation between actual investments and income, as well as the size of the finance industry, a model is definitely needed in order to recover the structural parameters of the economy. This is what I have tried to do in this paper.

This approach also offers a different perspective on the issue of credit constraints. The existing literature has provided much evidence on the role of financial frictions by looking at the investment behavior of individual firms. It is fair to say, however, that there is not much consensus regarding the quantitative importance of these frictions for the whole economy, because it is hard to aggregate the various studies into one meaningful number. I have taken a different perspective on the issue. In essence, I have asked: if financial frictions were not important, why would we spend 9% of GDP on financial services?

The analysis presented here is only a first step, however, mainly because it does not take into account the demand for financial services by households or the globalization of the finance industry. Incorporating these two forces is a task for future research.
References


Appendix

A Proof of proposition 4

Consider first the limit when \( \mu \to 0 \). First, rewrite (17) as

\[
\frac{\mu}{\pi m(\alpha_l)} = \frac{\bar{\alpha}}{\pi v} + 1 - F(\alpha_l) + \int_{\alpha_l}^{\alpha_h} \left( 1 - \frac{m(\alpha)}{m(\alpha_l)} \right) dF(\alpha).
\]

As \( \mu \) goes to zero, \( m(\alpha_l) \) also goes to zero and \( \alpha_l \to \alpha_h \). We need to evaluate the limit of the integral. For all \( \alpha \in [\alpha_l, \alpha_h] \), we know that \( 0 < m(\alpha) < m(\alpha_l) \) and therefore:

\[
\left| \int_{\alpha_l}^{\alpha_h} \left( 1 - \frac{m(\alpha)}{m(\alpha_l)} \right) dF(\alpha) \right| \leq F(\alpha_h) - F(\alpha_l).
\]

We can see that the integral goes to zero as \( \mu \) goes to zero and:

\[
\lim_{\mu \to 0} \frac{\mu}{\pi m(\alpha_l)} = \frac{\bar{\alpha}}{\pi v} + 1 - F(\alpha_h).
\]

Using equation (19), we see that \( \frac{1-n}{n} = \pi / \mu \int_{\alpha_l}^{\alpha_h} m(\alpha) dF(\alpha) \). Since the monitoring function \( m \) is decreasing in \( \alpha \), it follows that:

\[
\left| \int_{\alpha_l}^{\alpha_h} m(\alpha) dF(\alpha) \right| \leq m(\alpha_l) (F(\alpha_h) - F(\alpha_l)).
\]

Since we have shown that \( \mu/m(\alpha_l) \) has a finite limit, it follows that \( 1-n \to 0 \). In the other limit when \( \mu \to \infty \), the result is clear from equation (19) since the integral of the right-hand-side is bounded by \( \int_0^{\alpha_h} m(\alpha) dF(\alpha) \). QED.

B Comparative Statics

Let \( \Delta [\cdot] \) denote the differential of any function or variable of interest. To prove the various propositions, I will use the differential of equation (18):

\[
\Delta \left[ \frac{\mu}{\pi} \right] = \left( \Delta \left[ \frac{\bar{\alpha}}{\pi v} \right] - \Delta [F](\alpha_l) \right) m(\alpha_l) + \left( \frac{\bar{\alpha}}{\pi v} + 1 - F(\alpha_l) \right) \Delta [m] - \int_{\alpha_l}^{\alpha_h} \Delta [mdF],
\]

and of equation (19):

\[
\left( \int_{\alpha_l}^{\alpha_h} mdF + \frac{\mu}{\pi} \right) \frac{\Delta [n]}{n} = \frac{1-n}{n} \Delta \left[ \frac{\mu}{\pi} \right] - \int_{\alpha_l}^{\alpha_h} \Delta [mdF] + m(\alpha_l) f(\alpha_l) \Delta [\alpha_l].
\]

These formula hold because the boundary terms with \( \alpha_l \) cancel out and because \( m(\alpha_h) = 0 \). The monitoring function from equation (15) can be written as:

\[
m(\alpha) = z - \theta + (\rho + \delta)(1 - \alpha).
\]

The differential of this equation is:

\[
\Delta [m] = \Delta [z] - \Delta [\theta] - (\rho + \delta) \Delta [\alpha] + \Delta [\rho + \delta](1 - \alpha).
\]
B.1 Proof of proposition 5
Consider the impact of a change in $\mu$. From equation (18), we see that

$$\Delta \left[ \frac{\mu}{\pi} \right] = - (\rho + \delta) \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{l}) \right) \Delta [\alpha_{l}]$$

So it is clear that $\alpha_{h}$ decreases with $\mu$. The effect on the size of the banking sector, however, is ambiguous. From the monitoring market clearing (19), we see that

$$\left( \int_{\alpha_{l}}^{\alpha_{h}} m dF + \frac{\mu}{\pi} \right) \frac{\Delta [n]}{n} = \frac{1 - n}{n} \Delta \left[ \frac{\mu}{\pi} \right] + m(\alpha_{l}) f(\alpha_{l}) \Delta [\alpha_{l}]$$

The sign of the RHS clearly depends on the value of $n$.

B.2 Proof of proposition 6
For a change in $z$, we get

$$(\rho + \delta) \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{l}) \right) \Delta [\alpha_{l}] = \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{h}) \right) \Delta [z]$$

So an increase in $z$ increases $\alpha_{h}$. The effect on the size of the financial sector is ambiguous:

$$\frac{\Delta [n]}{n} \left( \int_{\alpha_{l}}^{\alpha_{h}} m (\alpha) dF (\alpha) + \frac{\mu}{\pi} \right) = m(\alpha_{l}) f(\alpha_{l}) \Delta [\alpha_{l}] - (F(\alpha_{h}) - F(\alpha_{l})) \Delta [z]$$

On the one hand, it takes more resources to monitor a given set of firms. On the other hand, the pool of firms that are monitoreed shrinks. Consider now a shift in the function $f$. We get

$$(\rho + \delta) \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{l}) \right) \Delta [\alpha_{l}] = - \Delta [F] \frac{m(\alpha_{l})}{n} - \int_{\alpha_{l}}^{\alpha_{h}} m (\alpha) \Delta [f] (\alpha) d\varepsilon$$

and

$$\left( \int_{\alpha_{l}}^{\alpha_{h}} m (\alpha) dF (\alpha) + \frac{\mu}{\pi} \right) \frac{\Delta [n]}{n} = - \int_{\alpha_{l}}^{\alpha_{h}} m (\alpha) \Delta [f] (\alpha) d\varepsilon + m(\alpha_{l}) f(\alpha_{l}) \Delta [\alpha_{l}]$$

B.3 Proof of proposition 7
Consider a shift in $\pi$

$$(\rho + \delta) \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{l}) \right) \Delta [\alpha_{l}] = \Delta \left[ \frac{1}{\pi} \right] \left( m(\alpha_{l}) \frac{\tilde{\alpha}}{\pi v} - \mu \right)$$

so if $\pi$ goes up, $\alpha_{l}$ goes up since $m(\alpha_{l}) \tilde{\alpha}/v < \mu$. The effect on $n$ is ambiguous:

$$\left( \int_{\alpha_{l}}^{\alpha_{h}} m (\alpha) dF (\alpha) + \frac{\mu}{\pi} \right) \frac{\Delta [n]}{n} = \frac{1 - n}{n} \Delta \left[ \frac{\mu}{\pi} \right] + m(\alpha_{l}) f(\alpha_{l}) \Delta [\alpha_{l}]$$

Consider a shift in $\theta$ note that it also means that $v$ goes up.

$$(\rho + \delta) \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{l}) \right) \Delta [\alpha_{l}] = m(\alpha_{l}) \Delta \left[ \frac{\tilde{\alpha}}{\pi v} \right] - \left( \frac{\tilde{\alpha}}{\pi v} + 1 - F(\alpha_{h}) \right) \Delta [\theta]$$

so clearly $\alpha_{l}$ goes down. The effect on $n$ is ambiguous.

$$\left( \int_{\alpha_{l}}^{\alpha_{h}} m (\alpha) dF (\alpha) + \frac{\mu}{\pi} \right) \frac{\Delta [n]}{n} = (F(\alpha_{h}) - F(\alpha_{l})) \Delta [\theta] + m(\alpha_{l}) f(\alpha_{l}) \Delta [\alpha_{l}]$$
### Table 1: Data

<table>
<thead>
<tr>
<th>Period</th>
<th>Finance Share of GDP</th>
<th>Investment Share of Firms with Income&lt;0.33*Capex</th>
<th>Credit Market Instruments over GDP</th>
<th>Total Corporate Liabilities over GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1955-1965</td>
<td>0.040</td>
<td>0.166</td>
<td>0.430</td>
<td>0.641</td>
</tr>
<tr>
<td>1966-1975</td>
<td>0.047</td>
<td>0.216</td>
<td>0.555</td>
<td>0.793</td>
</tr>
<tr>
<td>1976-1985</td>
<td>0.057</td>
<td>0.275</td>
<td>0.609</td>
<td>1.074</td>
</tr>
<tr>
<td>1986-1995</td>
<td>0.071</td>
<td>0.350</td>
<td>0.701</td>
<td>1.202</td>
</tr>
<tr>
<td>1996-2005</td>
<td>0.086</td>
<td>0.398</td>
<td>0.714</td>
<td>1.313</td>
</tr>
</tbody>
</table>

Notes: Finance Share of GDP is the value added of the Finance and Insurance industry over the value added of all private industries. Investment share of 'low cash' firms is the fraction of all capital expenditure in Compustat accounted for by firms whose income is less than a third of their capital expenditures. Credit Market Instrument over GDP includes all non financial corporate instruments outstanding. Source: National Income and Product Accounts, Compustat, and Flow of Funds.
<table>
<thead>
<tr>
<th></th>
<th>Finance Value Added over GDP</th>
<th>Fraction of Constrained Firms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting Value (from model)</td>
<td>0.0395</td>
<td>0.1324</td>
</tr>
<tr>
<td>Final Value (from model)</td>
<td>0.0780</td>
<td>0.1324</td>
</tr>
<tr>
<td>Demand Shift: predicted value without productivity gains in financial sector</td>
<td>0.0725</td>
<td>0.1957</td>
</tr>
<tr>
<td>Productivity Gains: predicted value without change in income-growth option correlation</td>
<td>0.0428</td>
<td>0.0962</td>
</tr>
</tbody>
</table>
Figure 1: The Size of the U.S. Financial Sector

Source: U.S. National Income and Product Accounts
Figure 2: No Moral Hazard
Figure 3: Moral Hazard Equilibrium

\[ 1 - F(z/\theta) \]

\[ \bar{\alpha} \]

\[ \pi \]

\[ e \]

\[ \theta - \delta - z \]

\[ \rho \]

\[ \theta - \delta \]

\[ r \]
Figure 4: Monitoring Equilibrium

Density function $f(\alpha)$

Constrained

$0 \leq \alpha_i \leq \alpha_h \leq 1$

Monitored Finance

Direct Finance

Self-Finance
Figure 5:
Effects of Productivity Gains in the Financial Sector

SIZE OF FINANCIAL SECTOR

FRACTION OF CONSTRAINED FIRMS

PRODUCTIVITY OF FINANCE, MU/MU0

PRODUCTIVITY OF FINANCE, MU/MU0
Figure 6: Simulation of Calibrated Model

Size of Financial Sector

- Actual
- Predicted
Figure 7: Simulation of Calibrated Model

Corporate Credit Market Instruments over GDP