Has the U.S. Finance Industry Become Less Efficient?

On the Theory and Measurement of Financial Intermediation

Thomas Philippon*

May 2012

Abstract

I provide a quantitative interpretation of financial intermediation in the U.S. over the past 130 years. Measuring separately the cost of intermediation and the production of financial services, I find that: (i) the quantity of intermediation varies a lot over time; (ii) intermediation is produced under constant returns to scale; (iii) the annual cost of intermediation is around 2% of outstanding assets; (iv) adjustments for borrowers' quality are quantitatively important; and (v) the unit cost of intermediation has increased over the past 30 years.

JEL: E2, G2, N2

*Stern School of Business, New York University; NBER and CEPR. This has been a very long project. The first draft dates back to 2007, with a focus on corporate finance, and without the long term historical evidence. This paper really owes a lot to other people, academics and non-academics alike. Darrell Duffie, Robert Lucas, Raghuram Rajan, Jose Scheinkman, Robert Shiller, Andrei Shleifer, and Richard Sylla have provided invaluable feedback at various stages of this project. Boyan Jovanovic, Peter Rousseau, Moritz Schularick, and Alan Taylor have shared their data and their insights, and I have greatly benefited from discussions with Lewis Alexander, Patrick Bolton, Markus Brunnermeier, John Cochrane, Douglas Diamond, John Geanakoplos, Gary Gorton, Robin Greenwood, Steve Kaplan, Anil Kashyap, Ashley Lester, Andrew Lo, Andrew Metrick, William Nordhaus, Matthew Rhodes-Kropf, David Robinson, Kenneth Rogoff, David Scharfstein, Hyun Shin, Jeremy Stein, Gillian Tett, Wallace Turbeville, and Luigi Zingales, as well as seminar participants at Stanford, Yale, NYU, Harvard, Chicago, Princeton, and the Paris School of Economics. I also thank Paul Krugman for his discussion at the 2011 NY Area Monetary conference, Axelle Ferrière, Peter Gross, Andrea Prestipino, Robert Turley, and Shaojun Zhang for research assistance, and the Smith Richardson Foundation for its financial support.

This paper is concerned with the theory and measurement of financial intermediation. Its contribution is to construct long time series on prices and quantities of intermediation, and to provide a quantitative interpretation of these series. Since the focus is on financial intermediation, the prices are spreads and fees earned by intermediaries, while the quantities are stocks and flows of financial assets and liabilities.

The role of the finance industry is to produce, trade and settle financial contracts that can be used to pool funds, share risks, transfer resources, produce information and provide incentives. Financial intermediaries are compensated for providing these services. The income received by these intermediaries measures the cost of financial intermediation. This income is the sum of all spreads and fees paid by non-financial agents to financial intermediaries, and it is also the sum of all profits and wages in the finance industry. The first contribution of the paper is empirical. I show that the income of financial intermediaries as a share of GDP varies a lot over time. The income share grows from 2% to 6% from 1870 to 1930. It shrinks to less than 4% in 1950, grows slowly to 5% in 1980, and then increases rapidly to more than 8% in 2010.

Given these large historical changes in the finance income share, it is natural to ask if there are commensurate changes in the production of financial services. The main contribution of the paper is to construct a theory-based measure of the "output" of the finance industry, and therefore, to obtain an estimate of the unit cost of financial intermediation. There two main motivations for undertaking such a project.

The first motivation is that the unit cost of intermediation is a critical parameter in many empirical and theoretical studies. Macro-finance models include, implicitly or explicitly, an equation that says that the borrowing rate is equal to the lending rate plus an intermediation cost: $r^{borrow} = r^{lend} + \psi$. Such an intermediation equation plays a central role in the literature that seeks to quantify the consequences of financial development for economic growth, while recent work has also focused on the macroeconomic consequences of a sudden increase in ψ, and on the link between ψ and intermediary capital, leverage and liquidity.[1] It is therefore crucial to know how large ψ is and whether it is stable over time.

The second motivation is to begin to quantify the efficiency gains from financial innovations. For instance, we would certainly like to know if the move from the traditional banking model towards the "originate-and-distribute" model has lowered the cost of funds for households and businesses. This is precisely what ψ should measure. It is important, however, to understand just how difficult the measurement problem is. In the traditional banking model, a bank would make a loan, keep it on its books, and earn a net interest income. This income would compensate for the cost of screening and monitoring the borrower, and then managing the duration and credit risk of the loan. In the originate and distribute model, the compensation for screening and monitoring shows up as a fee income, while the compensation for managing duration and credit risk might show up as trading profits in some hedge fund.

[1]For an analysis financial and economic development, see Greenwood, Sanchez, and Wang (2010), Buera, Kaboski, and Shin (2011), and Midrigan and Xu (2011). For an analysis of the macroeconomic consequences of negative shocks to financial intermediation, see Curdia and Woodford (2009), Gertler and Kiyotaki (2010), Hall (2011), Christiano and Ikeda (2011), Corsetti, Kuester, Meier, and Müller (2011) who all build on the classic contribution of Bernanke, Gertler, and Gilchrist (1999). For an analysis of liquidity, see Brunnermeier and Pedersen (2009), Gertler and Karadi (2011), He and Krishnamurthy (2012), and Moore (2011).

There would be no interest income and no direct measure of an interest rate spread. Similar issues immediately appear when we think about mutual funds, shadow banks, credit derivatives, etc.

Given these difficulties, my strategy is to take an aggregate view of the intermediation process and to construct a consistent, theory-based measure of output for the finance industry as a whole. I start by introducing financial services for firms and households in the neoclassical growth model. Conceptually, it is useful to distinguish three types of services:[2]

- (i) Liquidity (means of payments, cash management);
- (ii) Transfer of funds (pooling funds from savers, screening and monitoring borrowers);
- (iii) Information (price signals, advising on M&As).

Note that financial firms typically produce a bundle of such services. For instance, risk management uses all three types of activities.

Services of type (i) and (ii) typically involve the creation of various financial assets and liabilities. For the credit market, I measure separately the quantities borrowed by households, farms and non-farm (non-financial) businesses, and the government. The sizes of these markets vary significantly over time. For instance, the business credit market is relatively large in the 1920s, small in the 1960s and large again after 1980, although not as large as in the late 1920s. The most important trend in recent years is the increase in household debt. For the equity market, I measure the market value of outstanding stocks as well as the flows of initial and seasoned offerings. For liquidity I measure deposits, repurchase agreements, and money markets mutual funds. For advising fees I construct a measure of M&A activity.

The next step is to aggregate the various types of credit, equity issuances and liquid assets into one measure of the quantity of services produced by the financial sector for the non-financial sector. Essentially, output should be measured as a weighted average of the various quantities of assets, and the weights should reflect the relative intermediation requirements. Theory provides guidance on how to do this. I use the first order conditions of the model to interpret various micro-evidence on interest rates, prices, and fees. This allows me to estimate the weights and construct the output measure.

I can then divide the income of the finance industry by the estimated output measure to obtain a measure of unit cost. I find that this (annual) unit cost is around 2% and relatively stable over time. In other words, I estimate that it costs two cents per year to create and maintain one dollar of intermediated financial asset. I also find clear evidence that financial services are produced under constant returns to scale. For instance, from 1947 to 1973 (a period of stable growth without major financial crises), real income per-capita increases by 80% and real financial assets by 250%, but my estimate of the unit cost of intermediation remains remarkably constant.

[2]This classification is motivated by the mapping between theory and measurement discussed throughout the paper. It differs a little bit from that of Merton (1995). I do not attempt in this paper to measure the informativeness of prices. This issue is tackled by Bai, Philippon, and Savov (2011). See the discussion in Section 5.

The final contribution of the paper is to perform quality adjustments to the output series. In corporate finance, the mix of new and old ventures changes significantly over time. The 1920s and 1990s are times of entry by young and risky firms, while the 1960s appear relatively more stable. Jovanovic and Rousseau (2005) have shown that these patterns are related to waves of technological change. Similarly, for household finance we see that relatively poor households have gained access to financial markets in recent years. The challenge is to account for the fact that these borrowers require more intermediation (more screening or monitoring) per unit of credit extended. Once again I rely on theory to make the quality adjustments. These adjustments appear to be quantitatively important. In the 1990s for instance, the raw output measure is around three times GDP, while the adjusted measure is around four times GDP.

The adjusted unit cost is also more stable than the unadjusted one. However, even with the quality adjustment, I find that the unit cost of intermediation has increased since the mid 1970s and is now significantly higher than it was at the turn of the twentieth century. In other words, the finance industry that sustained the expansion of railroads, steel and chemical industries, and later the electricity and automobile revolutions seems to have been more efficient than the current finance industry. Surprisingly, the tremendous improvements in information technologies of the past 30 years have not led to a decrease in the average cost of intermediation, or at least not yet. One possible explanation for this puzzle is that improvements in information technology have been cancelled out by zero-sum activities, perhaps related to the large increase in secondary market trading.

Related literature

Financial intermediation does not have a benchmark quantitative framework in the way asset pricing does. By using a model to interpret long time series of prices and quantities, this paper shares the spirit of Mehra and Prescott (1985). It also articulates a puzzle for future research to solve. But because financial intermediation is a more heterogenous field than asset pricing, the paper builds on several strands of literature in finance and monetary economics,

The first strand is the theory of banking and intermediation. While stylized and focused on macroeconomic predictions, the model developed below is consistent with leading theories of financial intermediation, such as Diamond and Dybvig (1983), Diamond (1984), Gorton and Pennacchi (1990), Holmström and Tirole (1997), Diamond and Rajan (2001), and Kashyap, Rajan, and Stein (2002). Gorton and Winton (2003) provide a review of the literature on financial intermediation. The focus of this paper differs from that literature in several ways: (i) the measurement of the costs of intermediation; (ii) the simultaneous modeling of household and corporate finance; and (iii) the use of an equilibrium model to interpret the historical evidence.

There is a large literature on financial development, which I do not have room to discuss here, except to say that it tends to focus on cross-sectional comparisons of countries at relatively early stages of financial development in order to understand the impact of finance on economic growth (e.g. Rajan and Zingales (1998)) and the determinants of

financial development (e.g. La Porta, Lopez-de Silanes, Shleifer, and Vishny (1998), Guiso, Sapienza, and Zingales (2004)). The literature typically focuses on corporate finance (Greenwood, Sanchez, and Wang (2010), Buera, Kaboski, and Shin (2011), Midrigan and Xu (2011)).[3] This paper is more closely related to a recent branch of the literature that seeks to provide risk-adjusted measures of financial productivity (Haldane, Brennan, and Madouros (2010), Basu, Inklaar, and Wang (2011)) and that considers the possibility of inefficient financial development (Glode, Green, and Lowery (2010), Bolton, Santos, and Scheinkman (2011)). Philippon and Reshef (2007) share the historical perspective of this paper, but study only wages and human capital in the finance industry. Kaplan and Rauh (2010) also focus on compensation. The large historical changes in the finance share of GDP were first documented and discussed in Philippon (2008), but that paper only focused on corporate credit, which I now estimate to be less than half of the output of the finance industry.[4]

In its account of liquidity services provided by the finance industry, the paper is also related to the classic literature on money and banking. Lucas (2000) provides a benchmark analysis of money demand. Kiyotaki and Moore (2008) study the interaction of liquidity, asset prices and aggregate activity. A recent branch of this literature has focused on the rise of market-based intermediation, also called shadow banking. Pozsar, Adrian, Ashcraft, and Boesky (2010) document the structure of shadow banking. Gorton and Metrick (2012), Stein (2012), and Gennaioli, Shleifer, and Vishny (2011) emphasize the importance of investors demand for risk free assets. Gorton, Lewellen, and Metrick (2012) argue that much of the shadow banking activity (pooling and tranching) happens to satisfy this demand for risk free assets. I attempt to account for these activities by measuring shadow deposits, such as money market mutual funds and repurchase agreements. The rise of shadow banking also diminishes the relevance of the traditional literature focused on efficiency in banking. That literature did provide measures of productivity in banking (see Wang, Basu, and Fernald (2008) for a discussion), but it focused on net interest income, which is only about half of the income of today's large banks (see the numbers for JP Morgan in the appendix). An important point developed below is that it is difficult to break down the income earned by the finance industry into economically meaningful components.

Finally, it is important to emphasize some limitations of my analysis. First, it does not deal with financial crises and risk taking. For instance, my output series include all corporate borrowing by Telecom companies in the late 1990s and all subprime and home equity borrowing by households in the mid 2000s. In doing so, I never ask whether borrowing is appropriate or excessive, and I therefore miss the crucial insights of Reinhart and Rogoff (2009). Similarly, I consolidate the earnings of financial intermediaries without controlling for systemic risk taking.[5]

[3]My approach is complementary to this literature and uses many of its important insights. The difference is that I focus on the evolution of the entire U.S. finance industry. As a result, both theory and measurement must be expanded. For instance, following Beck, Demirguc-Kunt, and Levine (2011), the literature uses cross-country data on interest-rate spreads to estimate financing frictions, e.g., Greenwood, Sanchez, and Wang (2012). To study the US finance industry, it is important to recognize that non-interest income (fees, trading revenues, etc.) is now the dominant source of income for financial firms (even for banks: see JPMorgan's 2010 annual report for instance), that consumer credit is at least as important as corporate credit, and that the shadow banking's creation of safe assets is driven by investors' liquidity demand (all these points are discussed in details below).

[4]The paper did not consider household credit, and did not account for liquidity services, which have become important with the rise of the shadow banking system. Other technical issues are discussed in Philippon (2012a).

[5]See for instance Adrian and Shin (2008), Krishnamurthy (2009), and Acharya, Pedersen, Philippon, and Richardson (2009) for

Finally, I might be overstating the output of the finance industry because I do not adjust for the role of GSEs in the mortgage market, analyzed by Scharfstein and Sunderam (2011) among others.

The remainder of the paper is organized as follows. In Section 1, I construct my measure of the cost of financial intermediation. Section 2 presents the benchmark model of corporate and household finance to organize the discussion. Section 3 presents measures of output for the finance industry, and computes the unit cost of intermediation. Section 4 presents the quality adjustments. Section 5 discusses the role of information technology, price informativeness, financial derivatives, risk sharing, and trading. Section 6 concludes.

1 Income Share of Finance Industry

In this section, I present the first main empirical fact: the evolution of the total cost of financial intermediation in the US over the past 140 years. As argued in the introduction, there is no simple way to break down the income earned by the finance industry into economically meaningful components. For instance, insurance companies and pension funds perform credit analysis, fixed income trading provides liquidity to credit markets, and securitization severs the links between assets held and assets originated. From a historical perspective, these issues are compounded by regulatory changes in the range of activities that certain intermediaries can provide. Rather than imposing arbitrary interpretations on the data, I therefore focus on a consolidated measure of income, the sum of all interest and non interest income earned by all financial intermediaries, irrespective of whether they are classified as private equity funds, commercial banks, insurance companies, or anything else.

1.1 Raw Data

The paper uses a lot of data sources. To save space all the details regarding the construction of the series are provided in a separate online appendix. I focus on the following measure:

$$\phi = \frac{Finance\,Income}{Total\,Income}.$$

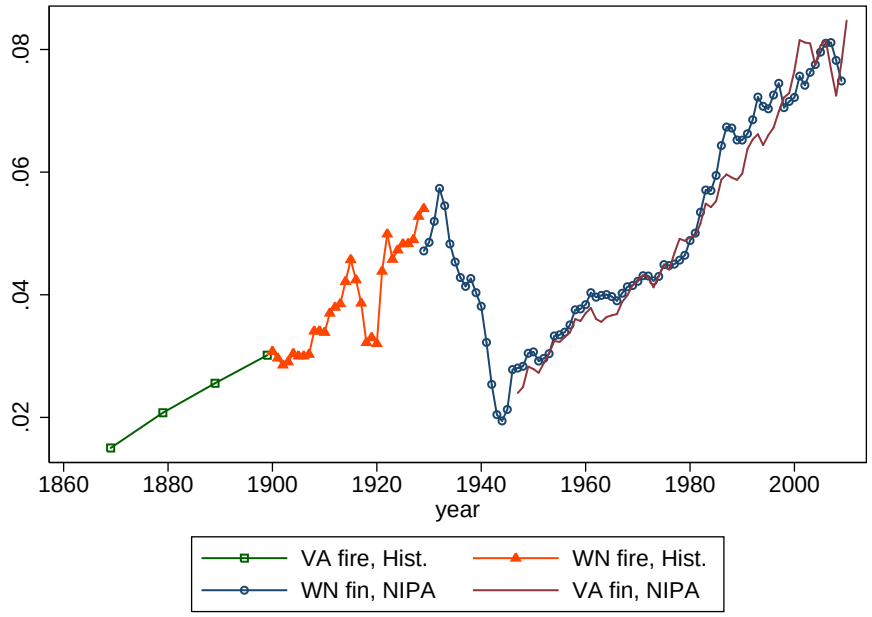
Conceptually, the best measure is total income (or value added), which is the sum of profits and wages. Whenever possible, I therefore construct ϕ as the GDP share of the finance industry, i.e., the nominal income of the finance industry divided by the nominal GDP of the U.S. economy. One issue, however, is that before 1945 profits are not always properly measured and value added is not available. In addition, it is sometimes difficult to measure the finance industry without imputed rents from the real estate sector. As an alternative measure I then use the labor compensation share of the finance industry, i.e., the compensation of all employees of the finance industry divided by the compensation of all employees in the U.S. economy.

recent discussions.

Figure 1 displays various measures of the share of the Finance and Insurance industry in the GDP of the United States estimated from 1870 to 2009. For the period 1947-2009, I use value added and compensation measures from the Annual Industry Accounts of the United States, published by the Bureau of Economic Analysis (BEA). For the post-war period, the two measures display the same trends. This means that, in the long run, the labor share in the finance industry is roughly the same as the labor share in the rest of the economy (in the short run, of course, profit rates can vary). For 1929-1947, I use the share of employee compensation because value added measures are either unavailable or unreliable. For 1870-1929 I use the Historical Statistics of the United States (Carter, Gartner, Haines, Olmstead, Sutch, and Wright, 2006).[6]

There are three important points to take away from Figure 1. First, the finance income share varies a lot over time. Second, the measures are qualitatively and quantitatively consistent. It is thus possible to create one long series simply by appending the older data to the newer ones. Third, finance as a share of GDP was smaller in 1980 than in 1925. Given the outstanding real growth over this period, it means that finance size is not simply driven by income per capita.

Figure 1: Income Share of Finance Industry



Notes: VA is value added, WN is compensation of employees, "fin" means finance and insurance, "fire" means finance, insurance, and real estate. For "NIPA", the data source is the BEA, and for "Hist" the source is the Historical Statistics of the United States.

[6]Other measures based on Martin (1939) and Kuznets (1941) give similar values. More details regarding the various data sources can be found in Philippon and Reshef (2007) and in the Data Appendix

1.2 Adjusted Measures

Before discussing theoretical interpretations it is useful to present adjusted series and consider the impact of globalization and the rise in services.

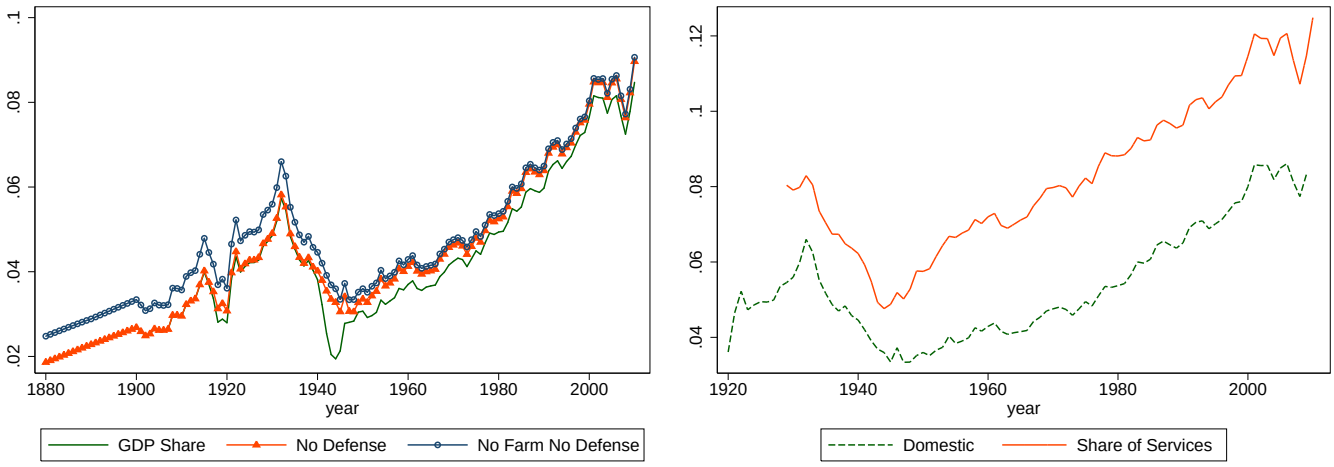
Wars

During peace time and without structural change, it would make sense to simply use GDP as the relevant measure of total income. Two factors can complicate the analysis, however. First, WWI and WWII take resources away from the normal production of goods and services. Financial intermediation should then be compared to the non-war related GDP. To do so, I construct a measure of GDP excluding defense spending. The second issue is the decline in farming. Since modern finance is related to trade and industrial development, it is also useful to estimate the share of finance in non-farm GDP.

The left panel of Figure 2 presents the finance share of non-defense GDP, and of non-farm, non-defense GDP (or compensation, as explained above). Both adjustments make the series more stationary. In particular, using non-defense GDP removes the spurious temporary drop in the unadjusted series during WWII.

I use the defense adjusted share as my main measure. The share of finance starts just below 2% in 1880. It reaches a first peak of almost 6% of GDP in 1932. Note that this peak occurs during the Great Depression, not in 1929. Between 1929 and 1932 nominal GDP shrinks, but the need to deal with rising default rates and to restructure corporate and household balance sheets keeps financiers busy. Similarly, the post-war peak occurs not in 2007 but in 2010, just below 9% of non-defense GDP.

Figure 2: Income Share of Finance (alternative measures)



Notes: GDP Share is the Income of the Finance Industry divided by GDP, constructed from the series in Figure 1. "No Defense" uses GDP minus defense spending, and "No Farm No Defense" uses non-farm GDP minus defense spending. Domestic shares excludes net exports of finance and insurance companies. Share of Services uses the BEA definition of services.

Other Services

Is finance different from other service industries? Yes. The right panel of Figure 2 also plots the share of finance in service GDP. It is mechanically higher than with total GDP, but the pattern is the same (the other fast growing service industry is health care, but it does not share the U-shaped evolution of Finance from 1927 to 2009).

Globalization

Figure 1 shows finance income divided by U.S. GDP. This might not be appropriate if financial firms export some of their services abroad. It turns out, however, that globalization does not account for the evolution of the finance income share. There are two ways to show this point.

The right panel of Figure 2 displays the ratio of *domestic* finance income to (non-defense) GDP. Domestic income is defined as income minus net exports of financial services. The figure is almost identical to the previous one. The reason is that the U.S., unlike the U.K. for instance, is not a large exporter of financial services. According to IMF statistics, in 2004, the U.K. financial services trade balance was +$37.4 billions while the U.S. balance was -$2.3 billions: the U.S. was actually a net importer. In 2005, the U.K. balance was +$34.9 billions, and the U.S. balance was +$1.1 billions. In all case, the adjustments are small.

The timing of globalization also cannot explain the evolution of the U.S. financial sector. Estevadeordal, Frantz, and Taylor (2003) show that the period 1870-1913 marks the birth of the first era of trade globalization (measured by the ratio of trade to output) and the period 1914-1939 its end. The period between 1918 and 1930, however, is the first large scale increase in the size of the finance industry, precisely as globalization recedes. For the more recent period, Obstfeld and Taylor (2002) and Bekaert, Harvey, and Lumsdaine (2002) show that financial globalization happens relatively late in the 1990s, while Figure 1 shows that the growth of the financial sector accelerates around 1980.

2 Benchmark Model

I now introduce a model of the economy with an explicit finance industry that provides services to households and businesses.[7] The model is useful for three reasons. First, the model provides comparative statics: which ratios should be constant on the balance growth path, what should be the impact of improvements in intermediation on these ratios, etc. Second the model provides guidance for constructing a well-defined measure of output for the finance industry: for instance, how should we compare deposits, residential mortgages, and corporate equity? Finally, an explicit model will be required to perform quality adjustment for heterogenous borrowers.

[7] It is critical to model financial services explicitly. It is well known that the properties of two-sectors models depend on the elasticity of substitution between the two sectors (Baumol, 1967). For instance, the nominal GDP share of sector i increases with relative technological progress in sector i if and only if the elasticity of substitution is less than one. In the context of financial intermediation, I will show that the elasticity depends both on the shape of the distribution of borrowers and on the efficiency of the supply of financial services. It is therefore not possible to take this elasticity as an (exogenous) parameter.

The model economy consists of households, a non-financial business sector, and a financial intermediation sector. The economy is non-stationary. The driving force is the labor-augmenting technological progress $A_t = (1+\gamma) A_{t-1}$. In the benchmark model borrowers are homogenous. This provides a tight characterization of equilibrium intermediation. I discuss heterogeneity and quality adjustments later.

2.1 Households

In the model, the finance industry provides two kinds of services to households: liquidity and credit, for transactions and consumption smoothing. Since household debt has an important life-cycle component (i.e., mortgages), I consider a setup with two types of households: some households are infinitely lived, the others belong to an overlapping generations structure.[8] Households in the model do not lend directly to one another. They lend to intermediaries, and intermediaries lend to firms and to other households.

Long-Lived Households

Long-lived households (index 0) own the capital stock and have no labor endowment. Liquidity services can be modeled using a cash-in-advance framework, or with money in the utility function. I use the later for simplicity, and I specify the utility function as $u(C_t, M_t) = \frac{(C_t M_t^{\nu})^{1-\rho}}{1-\rho}$, where C is consumption and M measures holdings of liquid assets. As argued by Lucas (2000), these homothetic preferences are consistent with the absence of trend in the ratio of real balances to income in U.S. data, and the constant relative risk aversion form is consistent with balanced growth. Let r be the interest rate received by savers. The budget constraint becomes $S_{t+1} + C_t + \psi_{m,t} M_t \leq (1+r_t) S_t$, where ψ_m is the price of liquidity services, and S are total savings.[9] The Euler equation $u_C(t) = \beta \mathbb{E}_t [(1+r_{t+1}) u_C(t+1)]$ can then be written as

$$M_t^{\nu(1-\rho)} C_{0,t}^{-\rho} = \beta \mathbb{E}_t \left[(1+r_{t+1}) M_{t+1}^{\nu(1-\rho)} C_{0,t+1}^{-\rho} \right].$$

The liquidity demand equation $u_M(t) = \psi_{m,t} u_C(t)$ is simply

$$\psi_{m,t} M_t = \nu C_{0,t}.$$

[8] The pure infinite horizon model and the pure OLG model are both inadequate. The infinite horizon model misses the importance of life-cycle borrowing and lending. The OLG model ignores bequests, and in the simple two-periods version households do not actually borrow: the young ones save, and the old ones eat their savings. The simplest way to capture all these relevant features is the mixed model. The standard interpretation is that long-lived households have bequest motives, and are therefore equivalent to infinitely lived agents. See also Mehra, Piguillem, and Prescott (2011) for a model where household save for retirement over an uncertain lifetime.

[9] See Lucas and Stokey (1987) and Sargent and Smith (2009) for a discussion of cash-in-advance models. Lucas (2000) uses the framework of Sidrauski (1967) with a more flexible functional form of the type $\left(C_t \varphi \left(\frac{M_t}{C} \right) \right)^{1-\rho}$. I use a Cobb-Douglass aggregator for simplicity given the complexity of the rest of the model. A more important difference with the classical literature on money demand is that I do not focus on inflation. Households save S at a gross return of $1+r$, while liquid assets yield $(1+r)/(1+\psi_m)$. So this model implies a constant spread between the lending rate and the rate on liquid assets. This is consistent with my interpretation of liquidity as not only money, but also money market funds shares and repurchase agreements.

Overlapping Generations

The other households live for two periods and are part of on overlapping generation structure. The young (index 1) have a labor endowment η_1 and the old (index 2) have a labor endowment η_2. We normalize the labor supply to one: $\eta_1+\eta_2=1$. The life-time utility of a young household is $u\left(C_{1,t},M_{1,t}\right)+\beta u\left(C_{2,t+1},M_{2,t+1}\right)$. I consider the case where they want to borrow when they are young (i.e., η_1 is small enough). In the first period, its budget constraint is $C_{1,t}+\psi_{m,t}M_{1,t}=\eta_1 W_{1,t}+\left(1-\psi_{c,t}\right)B_t^c$. The screening and monitoring cost is $\psi_{c,t}$ per unit of borrowing. In the second period, the household consumes $C_{2,t+1}+\psi_{m,t+1}M_{2,t+1}=\eta_2 W_{t+1}-\left(1+r_{t+1}\right)B_t^c$. The Euler equation for short-lived households is

$$\left(1-\psi_{c,t}\right)M_{1,t}^{\nu(1-\rho)}C_{1,t}^{-\rho}=\beta\mathbb{E}_t\left[\left(1+r_{t+1}\right)M_{2,t+1}^{\nu(1-\rho)}C_{2,t+1}^{-\rho}\right].$$

Their liquidity demand is identical to the one of long-lived households.

2.2 Non Financial Businesses

Non-financial output is produced with constant returns technology $Y_t=F\left(A_t n_t,K_t\right)$. The capital stock K_t depreciates at rate δ. Each unit of capital requires $1-\bar{x}$ units of screening and monitoring from intermediaries. Section 4 derives $\bar{x}$ endogenously from a standard moral hazard model, but for now I take it as a parameter. Let $\psi_{k,t}$ be the price of corporate financial intermediation. Non financial firms therefore solve the following program $\max_{n,K}F\left(A_t n,K\right)-\left(r_t+\delta+\left(1-\bar{x}\right)\psi_{k,t}\right)K-W_t n$. Capital demand equates the marginal product of capital to its user cost:

$$\frac{\partial F}{\partial K}\left(A_t n_t,K_t\right)=r_t+\delta+\left(1-\bar{x}\right)\psi_{k,t}. \tag{1}$$

Similarly, labor demand equates the marginal product of labor to the real wage:

$$A_t\frac{\partial F}{\partial n}\left(A_t n_t,K_t\right)=W_t. \tag{2}$$

Finally, I assume that the production function is Cobb-Douglass.[10]

Assumption: $F\left(A_t n_t,K_t\right)=\left(A_t n_t\right)^{\alpha}K_t^{1-\alpha}$

The key point to understand is that this model is fundamentally a user-cost model. The capital-output ratio is pinned down by equation (1), and improvements in finance lower the user cost of capital. For simplicity I describe intermediaries' services as screening and monitoring of primary issuances, but the model also captures business risk management. One well-understood benefit of risk management is to reduce the cost of financial distress, which is the

[10] Philippon (2012a) discusses the consequences of assuming different production function. The main issue is the elasticity of substitution between capital and labor, which is 1 under Cobb-Douglass technology. It turns out that this does not entail much loss of generality because qualitatively different results only happen for elasticity values above 6, which is far above the range of empirical estimates.

net present value of the deadweight losses incurred in states of the world where firms are in financial distress. These deadweight losses include bankruptcy costs and foregone investment opportunities (see Almeida and Philippon (2007) for a discussion). The standard way to model financial distress is to assume that it destroys a fraction of the firm's capital. The risk of future financial distress is then equivalent to a higher depreciation rate δ, and, conversely, risk management leads to a lower δ. Since the user cost is $r+\delta+(1-\bar{x})\,\psi_{k,t}$, it does not matter whether δ decreases or whether $\psi_{k,t}$ decreases. Therefore my framework will properly account for improvements in risk management.

2.3 Intermediation Equilibrium

There is a long tradition of modeling financial services. I do not attempt to do justice to this rich literature. Rather, I highlight the macroeconomic implication of technological progress in the finance industry on the size of credit markets and the GDP share of the industry.

Financial services are produced with capital and labor with constant returns to scale. Philippon (2012a) discusses in details the implications of various production functions, when financial intermediaries explicitly hire capital and labor. These issues are not central here, and I therefore assume financial services are produced from final goods with a marginal cost ζ_t. The quantity of financial services is given by

$$Y_t^{\phi} = \bar{\mu}_c B_t^c + \bar{\mu}_m M_t + \bar{\mu}_k B_t^k, \tag{3}$$

where $B_t^k = (1-\bar{x})\,K_t$ and $\bar{\mu}$ measures the required intermediation intensity. For instance, if it is more complicated to monitors entrepreneurs than households, we would have $\bar{\mu}_k > \bar{\mu}_c$. It is immediate from (3) that for each market $j = c, m, k$ we have

$$\psi_{j,t} = \zeta_t \bar{\mu}_j. \tag{4}$$

The parameters $\bar{\mu}$'s are convenient to describe various properties of the model. For instance, a decrease in $\bar{\mu}_k$ can be interpreted either as an improvement in the creditworthiness of businesses, or as an improvement specific to corporate finance intermediation. Finally, note that $\bar{\mu}_j$ is really the *average* intermediation intensity of market j. The benchmark model abstracts from heterogeneity within each market, but heterogeneity among firms and households will play a critical role when we study quality adjustments in Section 4.

An *equilibrium* in this economy is a sequence for the various prices and quantities listed above such that households choose optimal levels of credit and liquidity, financial and non financial firms maximize profits, and the labor and capital markets clear. This implies $n_t = 1$ and

$$S_t = K_{t+1} + B_t^c.$$

Aggregate GDP in this economy is defined by $\bar{Y}_t \equiv Y_t + \zeta Y_t^{\phi}$ and the finance share of GDP discussed in Section 1 is

$$\phi_t \equiv \frac{\zeta_t Y_t^{\phi}}{Y_t + \zeta_t Y_t^{\phi}}.$$

2.4 Balanced Growth

Let us now characterize an equilibrium with constant productivity growth in the non-financial sector (γ) and constant efficiency of intermediation (ζ). From now on, I use lower-case letters for de-trended variables, i.e. variables scaled by the current level of technology. For instance, for capital I write $k \equiv \frac{K_t}{A_t}$, for consumption of agent i $c_i \equiv \frac{C_{i,t}}{A_t}$, and for the productivity adjusted wage: $w \equiv W_t/A_t$. Let us discuss the main features of the equilibrium. On the balanced growth path, M grows at the same rate as C. The Euler equation for long-lived households becomes $1 = \beta \mathbb{E}_t \left[(1 + r_{t+1}) \left(\frac{C_{t+1}}{C_t} \right)^{\nu(1-\rho)-\rho} \right]$, so the equilibrium interest rate is simply pinned down by

$$\beta (1 + r) = (1 + \gamma)^{\theta}. \tag{5}$$

where $\theta \equiv \rho - \nu (1 - \rho)$. With Cobb-Douglass technology, the capital labor ratio is

$$\frac{k}{n} = \left(\frac{1 - \alpha}{r + \delta + (1 - \bar{x}) \psi_k} \right)^{\frac{1}{\alpha}}.$$

Since $n = 1$ in equilibrium, this is also the aggregate stock of capital. Non financial GDP is $y = k^{1-\alpha}$. The real wage is

$$w = \alpha k^{1-\alpha} = \alpha y.$$

The wage is increasing in the efficiency of corporate finance. Note that the capital stock and corporate intermediation are independent of consumer credit and liquidity. The fact that consumer credit does not crowd out corporate credit is not a general property, it comes from the simplifying assumption that finance uses no labor. In the general case, there is crowding out (see Philippon (2012a)).

Given the interest rate in (5), the Euler equation of short lived households is simply

$$c_1 = (1 - \psi_c)^{\frac{1}{\theta}} c_2. \tag{6}$$

In addition, we have $\psi_m m = \nu c$ for each cohort. The budget constraints are therefore $(1 + \nu) c_1 = \eta_1 w + (1 - \psi_c) b$ and $(1 + \nu) c_2 = \eta_2 w - \frac{1+r}{1+\gamma} b$. We can then use the Euler equations and budget constraints to compute the borrowing of young households

$$b_c = \frac{(1 - \psi_c)^{\frac{1}{\theta}} \eta_2 - \eta_1}{1 - \psi_c + (1 - \psi_c)^{\frac{1}{\theta}} \frac{1+r}{1+\gamma}} w. \tag{7}$$

If ψ_c is 0, we have perfect consumption smoothing: $c_1 = c_2$. From the perspective of current consumption, borrowing

costs act as a tax on future labor income. If ψ_c is too high, no borrowing takes place and the consumer credit market collapses. The bigger the ratio η_2/η_1 the larger the borrowing. For instance, increased years schooling generates more borrowing, and a larger financial sector. Improvements in corporate finance increase household debt b_c because they increase w, but b_c/y remain constant since $w = \alpha y$.

The quantity of services produced of the finance industry is

$$y^\phi = \bar{\mu}_c b_c + \bar{\mu}_m m + \bar{\mu}_k b_k.$$

We already know b_c and b_k, and liquidity demand is

$$m = \frac{\nu c}{\psi_m}.$$

Since $\psi_m = \zeta \mu_m$, so we only need to measure aggregate consumption as

$$c = \frac{1}{1+\nu}\left(w - \psi_c b_c + (r-\gamma)\, k\right).$$

Improvements in corporate finance increase liquidity demand because they increase the consumption output ratio. When ψ_k goes down, k/y goes up while b/y is unchanged, therefore $\nu c_0/y$ goes up.

The dollar value of the finance industry's output is $\zeta y^\phi = \psi_c b_c + \psi_m m + \psi_k b_k$ and the finance share of GDP is

$$\phi = \frac{\zeta y^\phi}{y + \zeta y^\phi}.$$

One important point is that the model does not predict an income effect, i.e., just because a country becomes richer does not mean that it should spend a higher fraction of its income on financial services. This is consistent with the fact that the finance share of GDP in Figure 1 is the same in 1980 as in 1925.[11] The following proposition summarizes the predictions of the theory.

Proposition 1. *There is a unique balanced growth path where the finance share of GDP ϕ, the unit cost of financial intermediation p^ϕ, and the financial ratios M/Y, B^c/Y and K/Y are constant. The equilibrium has the following features*

(i) Improvements in corporate finance increase output, the real wage, and the capital-output ratio, household debt increases proportionally to GDP, and liquidity increases more than proportionally;

(ii) Improvements in household finance increase household debt, consumption and liquidity, but do not affect the real wage;

[11] Similarly, Bickenbach, Bode, Dohse, Hanley, and Schweickert (2009) show that the income share of finance has remained remarkably constant in Germany over the past 30 years. More precisely, using KLEMS for Europe (see O'Mahony and Timmer (2009)) one can see that the finance share in Germany was 4.3% in 1980, 4.68% in 1990, 4.19% in 2000, and 4.47% in 2006.

(iii) Improvements in liquidity management increase consumption, but do not affect household debt or the real wage;

(iv) Demand shifts (e.g. η_2) increase the finance income share while supply shifts (ζ) have an ambiguous impact on the income share.

Proof. See Appendix. □

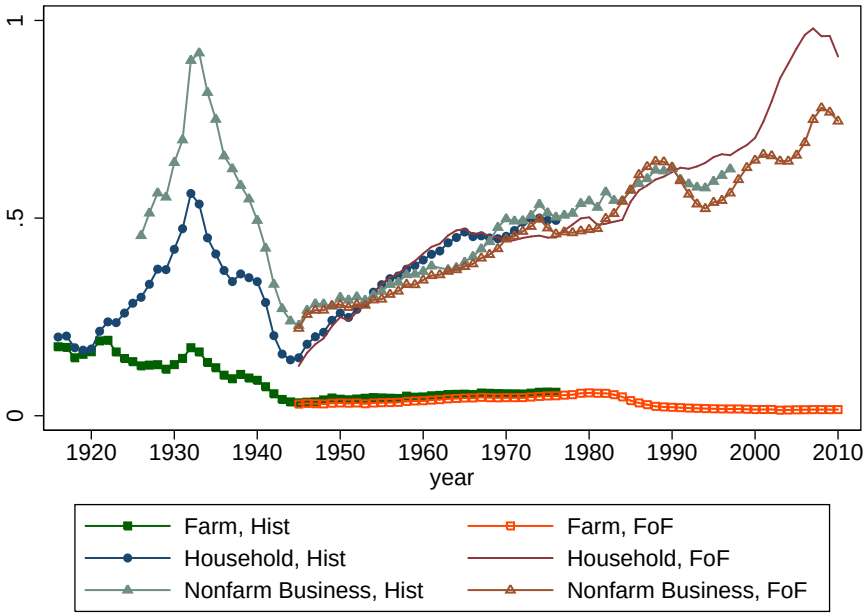
3 Output and Unit Cost

In this section I construct empirical proxies for $\frac{m}{y}; \frac{b_c}{y}; \frac{b_k}{y}$ as well as other elements of what the finance industry does. Note that financial derivatives are discussed in Section 5.

3.1 Credit Markets

Figure 3 presents credit liabilities of farms, households and the business sector (corporate and non-corporate). The first point to take away is the good match between the various sources. As with the income share above, this allows us to extend the series in the past. Two features stand out. First, the non-financial business credit market is not as deep even today as it was in the 1920s. Second, household debt has grown significantly over the post-war period.[12]

Figure 3: Debt over GDP



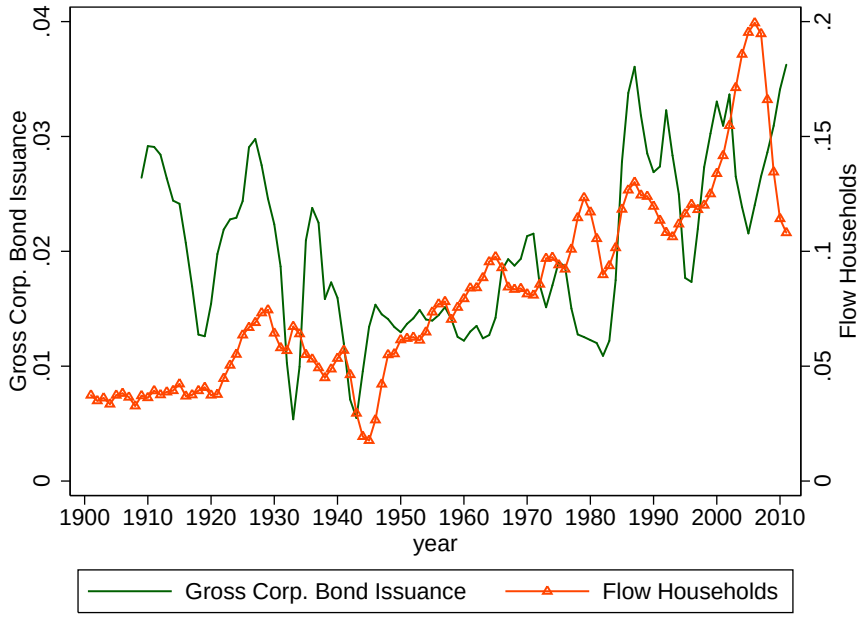
Notes: "FoF" is flow of funds, "Hist" is Historical Statistics of the United States. Business includes non-farm corporate and non-corporate debt.

[12] I have also constructed credit liabilities of financial firms. Financial firms have recently become major issuers of debt. Banks used to fund themselves with deposits and equity, and almost no long term debt. Today they issue a lot of long term debt. Note that it is critical to separate financial and non-financial issuers. What should count as output for the finance industry are only issuances by non-financial firms.

To extend the credit series before 1920, I use data on home mortgages provided by Schularick and Taylor (forthcoming). I also use the balance sheets of financial firms. I measure assets on the balance sheets of commercial banks, mutual banks, savings and loans, federal reserve banks, brokers, and life insurance companies. I define total assets as the sum of assets of all these financial firms over GDP. I use this series to extend the total non-financial debt series (households & non corporates, farms, corporates, government). I regress total credit on total assets and use the predicted value to extend the credit series.

In the theory outlined earlier, there is no distinction between outstanding assets and new issuances. In the data the two can be different and it is useful to consider stocks and flows separately. Figure 4 shows the issuances of corporate bonds by non-financial corporations as well as a measure of household credit flows.[13] Note that issuances collapse in the 1930s when the debt to GDP ratio peaks, in part because of deflation. There is thus a difference of timing between measures of output based on flows (issuances) versus levels (outstanding). Figure 4 also shows a measure of household debt issuance.

Figure 4: Debt Flows



Notes: Gross Issuance of Corporate Bonds is a three-year centered moving average of gross issuances of bonds by non-financial firms, from Baker and Wurgler (2000). Household Issuance is based on the Flow of Funds and the Historical Statistics of the United States and

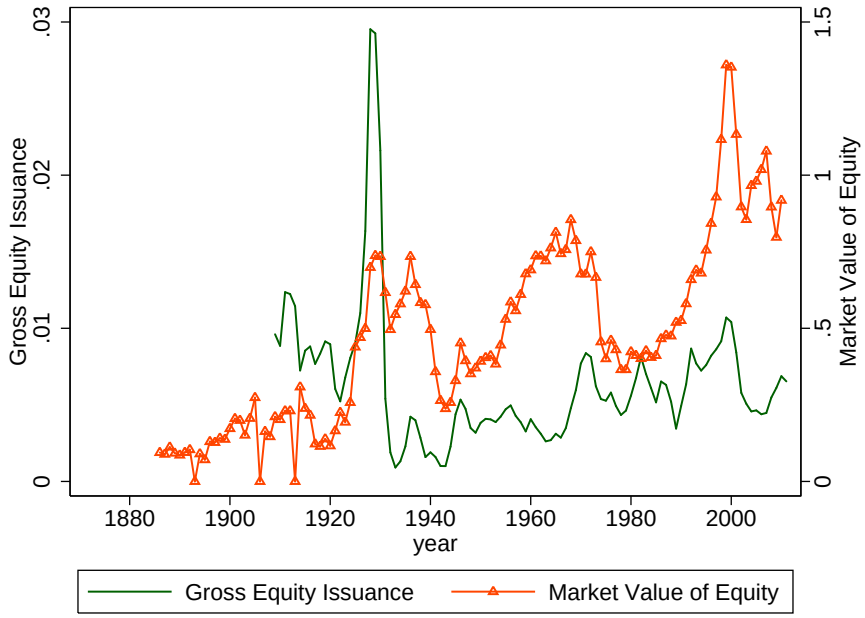
3.2 Equity Market

The equity market is difficult to deal with because of valuation effects. Stocks, unlike bonds, are recorded at market value. The ratio of the market value of equity to GDP can fluctuate without any intermediation services (i.e., without any issuance of equity). Another problem is that net issuances are often negative. However, negative net

[13] When I do not have a separate measure of flows, I assume a runoff rate consistent with the average ratio of flow to level, and I create the flow measure from the level series. Details are in the data appendix.

issuances do not imply that no intermediation services are produced. To deal with these problems I use three measures of equity production: total market value over GDP, IPO proceeds over GDP, and gross (non-financial) equity offerings over GDP. The advantage of using IPOs is that they provide a good measure of entry and growth by young firms, whose screening and monitoring requirements are certainly higher than those of established companies. Thus, the IPO series will allow me to control for heterogeneity. Quality-adjustments based on heterogeneity among borrowers are presented in the next section.

Figure 5: Equity Value and Gross Issuance over GDP



Notes: Market Value of non-financial corporate firms from the Flow of Funds and from CRSP. Gross Equity Issuance is a three-year centered moving average of gross issuances of stocks by non-financial firms, from Baker and Wurgler (2000)

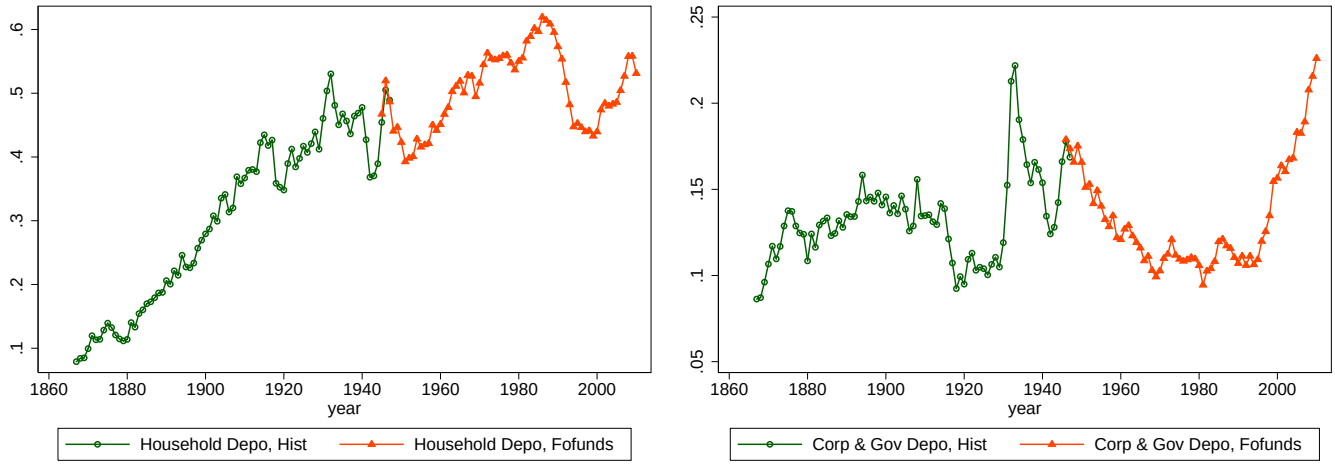
3.3 Money and Liquidity

In addition to credit (on the asset side of banks), households, firms and local governments benefit from payment and liquidity services (on the liability side of banks and money market funds). For households, I use the total currency and deposits, including money market fund shares, held by households and nonprofit organizations. The left panel of Figure 6 shows the evolution of this variable.

An important issue in the measurement of liquidity provision is the rise of the shadow banking system. Gorton, Lewellen, and Metrick (2012) argue that a significant share recent activities in the financial sector was aimed at creating risk free assets with money-like features. For firms (incorporated or not), I follow Gorton, Lewellen, and Metrick (2012) and I treat repos as shadow deposits. The series is thus the sum of checkable deposits and currency, time and savings deposits, money markets mutual funds shares, and repos (by non financial firms).[14]

[14]I have experimented with an adjustment for the fact that deposit insurance provided by the government makes it cheaper for private agents to create deposits. The adjustments seem rather arbitrary and did not make a significant difference so I dropped it. But more

Figure 6: Liquidity: Depos and Repos



Notes: Deposits, Repurchase Agreements, and Money Market Mutual Funds shares over GDP. Sources are Historical Statistics of the United States and Flow of Funds.

3.4 Aggregation

The model presented in Section 2 suggests measuring the output of the finance industry as the weighted sum of assets intermediated. More precisely, we have $y^{\phi} = \bar{\mu}_c b_c + \bar{\mu}_m m + \bar{\mu}_k b_k + \bar{\mu}_g b_g$, where b_k is corporate intermediation, b_c is household borrowing, m are deposits, and b_g is government debt. The main issue is that, as explained earlier, there is no satisfactory way to link a particular income to a particular activity, especially over long periods of time. This precludes a direct estimation of the $\bar{\mu}$'s.[15] It is possible, however, to obtain indirect estimates by using the first order conditions of the model together with microeconomic evidence on the prices of various financial services. I proceed in two steps. I first assume that the $\bar{\mu}$'s are constant over time, and in Section 4, I estimate time varying quality adjustments to the $\bar{\mu}$'s.

Households save S at a gross return of $1 + r$, while liquid assets yield $(1 + r) / (1 + \psi_m)$. In the first order condition for liquidity demand, ψ_m is the price that investors are willing to pay for liquidity, and it can be measured as an opportunity cost to savers. Table 1 presents the relevant rates and returns. Over the period 2002-2011, the Vanguard Short Term Treasury fund has returned 3.65%, with an expense ratio of 0.22.[16] This is the opportunity cost of cash, but cash is a relatively small fraction of liquid assets. Over the same period, the Vanguard Prime Money Market fund has returned 1.82% with an expense ratio of 0.20. The difference in returns between these two essentially risk free instruments is 1.8%. One can also interpret ψ_m as the cost of creating liquid assets. This cost can be charged as a redemption fee for investors. This fee is around 2% and is consistent with the trading costs incurred by mutual funds upon withdrawals (see Chen, Goldstein, and Jiang (2010) for a discussion). Based on this

quantitative work would clearly be needed here.

[15] This is not a simple accounting issue. Financial tasks are deeply intertwined. Insurance companies and pension funds perform their own independent credit analysis. Banks act as market makers. Investment banks behave as hedge funds. In addition, the mapping from industry to tasks has changed over time with the development of the originate and distribute model in banking.

[16] Vanguard data accessed on March 11, 2012.

Table 1: U.S. Interest Rates and Returns, 2002-2011

Variable	Value (%)
Average 3M Treasury Bills	1.79
Average 1M Certificate of Deposits	2.14
Average 1Y Gov. Bond	2.10
Average 10Y Gov. Bond	3.95
Average Aaa Corporate	5.47
Average Prime Bank Loan	5.02
Average 30Y Conventional Mortgage	5.71
Average Baa Corporate	6.64
Vanguard Prime Money Market Fund Return	1.82
Vanguard Short Term Treasury Fund Return	3.65

Source: FRED, and Vanguard. All values are average over 2002-2011.

discussion, I will therefore assume that $\psi_m = 2\%$.

Next I need to compare the intermediation requirements of households and businesses. Corporate intermediation itself is made of equity and debt financing, hence $\bar{\mu}_k b_k = \bar{\mu}_k^d d_k + \bar{\mu}_k^e e_k$, where e_k is corporate equity and d_k is corporate debt. For corporate issuances, Altinkilic and Hansen (2000) report fees of 3% to 4% for equity and about 1% for bonds. For flows, I therefore assume $\bar{\mu}_k^e = 3.5\bar{\mu}_k^d$. These numbers seem to be in line with recent reports by large investment banks. For instance, JP Morgan's 2010 annual report suggests underwriting fees around 0.70% for debt, and around 2.46% for equity (see Appendix). For households we can look at the mortgage market. Sirmans and Benjamin (1990) report fees of 0.50% to 0.70%. For other types of consumer credit these fees are certainly larger. Table 1 suggests that once issued, high quality corporate and consumer debt trades at similar prices. I therefore set $\bar{\mu}_k^d = \bar{\mu}_c$ so that it is equally difficult to extend credit to firms or to households.

I now need to construct two series for each type of output, one for the flow of new assets, and one for the level of outstanding assets.[17] Before doing so, however, I need to discuss M&A activities.

M&As

An important activity of financial intermediaries is advising on mergers and acquisitions. Scharfstein (1988) argues that the threat of takeover can improve managerial incentives. Rhodes-Kropf and Robinson (2008) show that M&As differ from other types of investment and require specific search efforts. From 1980 to 2010, I use data from SDC and Bloomberg to compute the value of merger deals. I then use historical data from Jovanovic and Rousseau (2005) to extend the series back to 1890. The next step is to apply the proper weight to the M&A series. M&A fees typically range from 1% for large deals to 4% for smaller ones. I assume that merger fees are 2% of the volume. This assumption is probably a bit higher than the weighted average fee, but there are also probably some ancillary activities associated with mergers and for which the finance industry is compensated.

[17]Note that both measures are relevant in theory. Screening models apply more directly to the flow of new issuances, while trading and monitoring models apply to both the flows and the levels.

Level and Flow Measures

The discussion so far suggests $\bar{\mu}_k^d = \bar{\mu}_c$ and $\psi_m = 2\%$. It turns out that the average cost of intermediation is also around 2%. This then implies that $\bar{\mu}_m = \bar{\mu}_c$. Anticipating this, and for ease of exposition, I therefore simply assume that $\bar{\mu}_m = \bar{\mu}_c$ (and this will in fact be consistent).

I also need to take into account the debt of the government. The issue is which weight to apply. On the one hand, government debt is risk-free and liquid, and it might actually help the functioning of financial markets (Krishnamurthy and Vissing-Jorgensen (2010), Greenwood, Hanson, and Stein (2011)). One option would then be to ignore government debt, i.e., to assume that it requires no intermediation services. But long term debt carries duration risk, and it needs to be traded, so intermediation requirements are not exactly zero. To be conservative, I assume that government debt intermediation requirements are 1/10 of that of private debt: $\frac{\bar{\mu}_g}{\bar{\mu}_c} = 0.1$.

Finally, for the level measure, I treat equity and debt similarly. The level measure is therefore:

$$y_{level}^{\phi} = b_c^{level} + d_k^{level} + e_k^{level} + m^{level} + 0.1 b_g^{level} + y_{M\&A}^{level}.$$

Flows correspond to issuances. I calibrate the relative weights using underwriter fees as explained earlier. The flow measure is therefore:

$$y_{flow}^{\phi} = b_c^{flow} + d_k^{flow} + 3.5 e_k^{flow} + m^{flow} + 0.1 b_g^{flow} + y_{M\&A}^{flow}$$

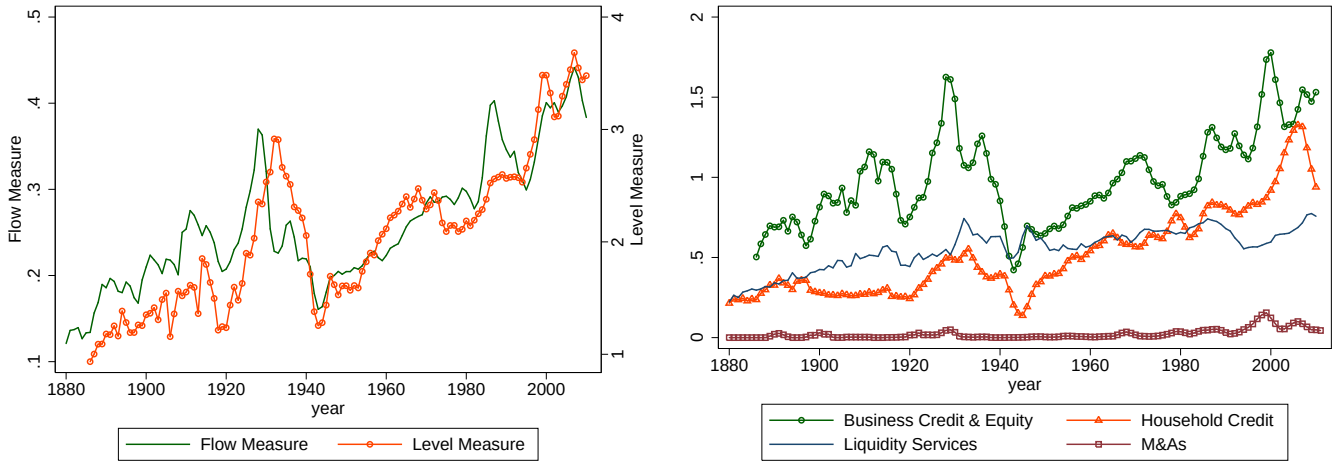
The two series are displayed in the left panel of Figure 7. The production of financial services increases steadily until WWI, and rapidly from 1919 to 1929. It collapses during the great depression and WWII. The flow and level measures share the same trends, but there are clear differences at medium frequencies. The flow measure is more stationary than the level measure. The flow measure collapses quickly during the great depression while the level measure peaks later in 1932-33 (this is exacerbated by deflation). A similar pattern seems to emerge during and after the great recession of 2008-2009.

Composite measures

As explained above, I have constructed two output series for the finance industry. One using the flows (gross issuances over GDP) and one using the levels (debt and equity over GDP). I now wish to combine the flow and level measures. On average, in the post-1950 period (where the data is most reliable), the ratio of the flow measure to the level measure is .118. In other words, the level measure is about 8.5 times the flow one. I therefore set $\lambda = \bar{y}_{flow}^{\phi} / \bar{y}_{level}^{\phi} = 0.118$ and I construct the composite measure as

$$y^{\phi} = \frac{1}{2}\left(y_{level}^{\phi} + \frac{y_{flow}^{\phi}}{\lambda} \right) \tag{8}$$

Figure 7: Output of Finance Industry



Notes. Left: aggregate measures of output for US finance industry, levels and flows, as shares of GDP. Right: composite measures (average of levels and flows) across broad functions.

There is no theoretical reason to prefer one measure over the other, so it seems logical to put equal weights on both. In any case, since the flow and level measures share the same trend, changing the weight changes the short run behavior of the composite measure, but not its long run behavior. Equation (8) implies that the scale of the composite measure is comparable to the scale of the level measure, so that one can interpret the unit cost of intermediation as an annual interest rate spread.

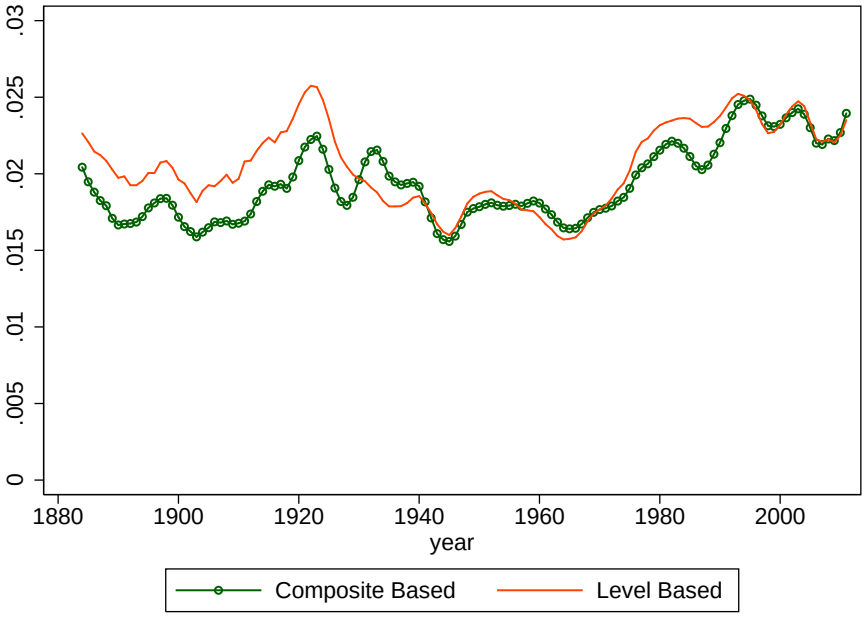
The right panel of Figure 7 presents the composite measures corresponding to 4 broad functions discussed earlier: credit and equity intermediation services to firms (farms, corporate, non-corporate), credit intermediation services to households, liquidity services to both, and M&A activities. In each case, the composite measure is based on the stock and flow measures aggregated as in (8). Note that the liquidity and M&A measures are not in fact composite measures since I do not have an independent flow measure for deposits or and independent level measure for M&As. It is clear from Figure 7 that credit intermediation for firms and households are the most volatile series. There is also a significant increase in liquidity services in the 2000s. M&As play a significant role mostly in the 1990s.

3.5 Unadjusted Unit Cost

Figure 8 estimates the cost of financial intermediation, defined as income divided by output. For income, I use domestic income, i.e., income minus net exports, as explained in Section 1. For output, I use the composite measure of equation (8), and also, as a robustness check, the simple level measure. There are two main points to take away. The first is that the ratio is relatively stable. Recall that we started from a series in Figure 1 that fluctuates by a factor of 5 (9% relative to less than 2%). All the debt, deposit and equity series also vary a lot over time. But their ratio, properly scaled, is quite stable. On Figure 8 it stays roughly between 1.5% and 2.5% over 130 years. One must also keep in mind that the model ignores business cycles and assumes constant real interest rates.

The second main point is that the finance cost index has been trending upward since the 1970s. A possible explanation is that the characteristics of borrowers have changed over time. In the next section, I show how quality adjustments go some way towards solving this puzzle. Before doing so, however, I present evidence of constant returns to scale in financial intermediation.

Figure 8: Unadjusted Unit Cost



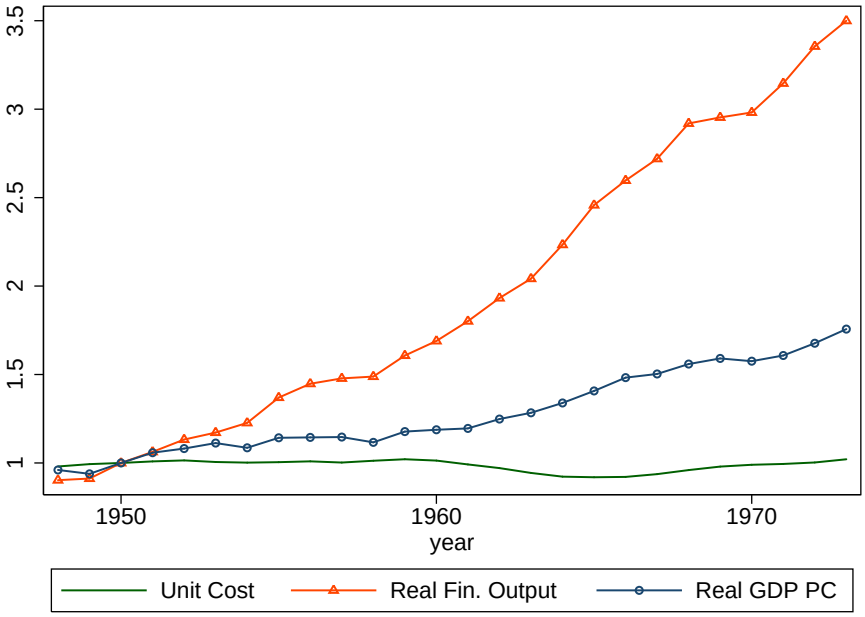
Notes: Total intermediation costs divided by production of credit assets, equity and deposits. Composite measure uses average of levels and flows.

3.6 Evidence of Constant Returns to Scale

An important assumption of the model is that financial services are produced under constant returns to scale. Figure 9 presents evidence consistent with this assumption. It uses the period 1947-1973, for two reasons. First, the post-war data is the most reliable, and stopping in 1973 allows me to exclude major oil shocks, inflation and other factors that might create short term noise in my estimates. Second, as I will discuss shortly, quality adjustments are not important over this period. Since these adjustments are difficult to implement, it is more convincing to first present the evidence without them.

From 1947 to 1973, real GDP per-capita increases by 80% and real financial assets by 250% (the finance output measured in constant dollars), but my estimate of the unit cost of intermediation remains fairly constant (all series are presented as ratios to their values in 1950). By 1973 people are a lot richer, financial markets are a lot larger, but the unit cost is the same. This provides clear evidence that the production of financial services has constant returns to scale.

Figure 9: Constant Returns to Scale



Notes: Unit cost of financial intermediation, composite measure of finance output, and real GDP per capita. Series normalized to one in 1950.

4 Quality Adjustments

The quantities of intermediation should be adjusted for the difficulty of monitoring/screening borrowers. Suppose that, for some reason, borrowers become harder to screen. A given amount of lending requires more intermediation. The income share of finance increases. Without proper quality adjustment, the raw measure would register a spurious increase in intermediation cost (i.e., a spurious decrease in efficiency of the finance industry). I now present quality adjustments to business borrowing and to household borrowing.

4.1 Corporate Finance

The homogenous borrower model described above is a useful benchmark, but it fails to capture some important features of corporate finance. To give just one example, corporate finance involves issuing commercial paper for blue chip companies as well as raising equity for high-technology start-ups. The monitoring requirements per dollar intermediated are clearly different in these two activities. More generally, my benchmark measure of business intermediation is an average of credit flows with different intermediation intensities. Measurement problems arise when the mix of high- and low-quality borrowers changes over time. Constant heterogeneity does not pose a problem: it amounts to a simple rescaling of the unit cost in Figure 8. Changes in the share of hard-to-monitor projects, however, present a challenge.

Let us therefore consider a model with heterogeneity and decreasing returns at the firm level.[18] Let k_t be the

[18] Decreasing returns in production are required to make room for heterogeneity since with constant returns borrowers that have even a slight financial disadvantage would not be able to enter.

Table 2: Summary of Models

Benchmark Model (Section 2)		
User Cost	ψ_k	$\frac{\partial F}{\partial K} = r + \delta + (1 - \bar{x})\,\psi_k$
Intermediation Intensity	$\bar{\mu}_k$	exogenous
Unit Cost	ζ	$\psi_k = \zeta \bar{\mu}_k$
Intermediated Credit	b_k	$b_k = (1 - \bar{x})\,k$
Real Intermediation	y_k^ϕ	$y_k^\phi = \bar{\mu}_k b_k$
Endogenous Monitoring Model		
Monitoring requirement	μ_k^j	$\mu_k^j = r + \delta - \pi(w) + (1 + r)\left(\xi_k - x^j\right)$
Real Intermediation	y_k^ϕ	$y_k^\phi = k_l \mu_k^l + k_h \mu_k^h$
Intermediation Intensity	$\bar{\mu}_k$	$\bar{\mu}_k = \frac{y_k^\phi}{b_k}$
Share of Low Cash Firms	s_k	$s_k = \frac{k_l}{k_l + k_h}$

(endogenous) number of firms, and let n_t be employment per-firm (so aggregate employment is $\bar{n}_t = k_t n_t$). Each firm needs exactly A_t units of capital and aggregate capital is $K_t = A_t k_t$. If it hires n workers it produces $A_t f(n)$ units of output, where f is increasing and concave. As before, we can scale all the variables by A_t to obtain a stationary model where k and n are constant over time. Firms choose employment to maximize (detrended) net income $\pi(w) \equiv \max_n f(n) - w_t n$. The labor demand schedule is the decreasing function $n(w)$ implied by the optimality condition $f'(n) = w$.

To capture financial intermediation in a tractable way, I assume that capital can be diverted and that the interests of the firm (or of the entrepreneur running it) are aligned with those of residual claimants who own a share x of the capital stock. In practice we can interpret x as inside equity or as retained earnings. I consider a model with two types of firms l and h, with $x^l < x^h$, and I study equilibria with free entry of l-firms and a given number, k_h, of h-firms. The Appendix describes the details of moral hazard and endogenous monitoring. For simplicity, I interpret x as cash on hand, and I refer to l-firms as low cash firms. The key point is that the model delivers the following monitoring demand function

$$\mu(x) = r + \delta - \pi(w) + (1 + r)(\xi - x),$$

where ξ measures the degree of diversion that would prevail without monitoring. The function $\mu(x)$ measures the quantity of intermediation services required for a firm with cash on hand x (and therefore external financing needs of $1 - x$). Firms with high values of x require less monitoring than firms with low values of x.

Table 2 summarizes the key variables and how they are related. The first half of the table refers to the benchmark model of Section 2, and the second half to the extended model with heterogenous borrowers presented in the Appendix.

Low cash firms mechanically need to raise more external finance than high cash firms. The important point is

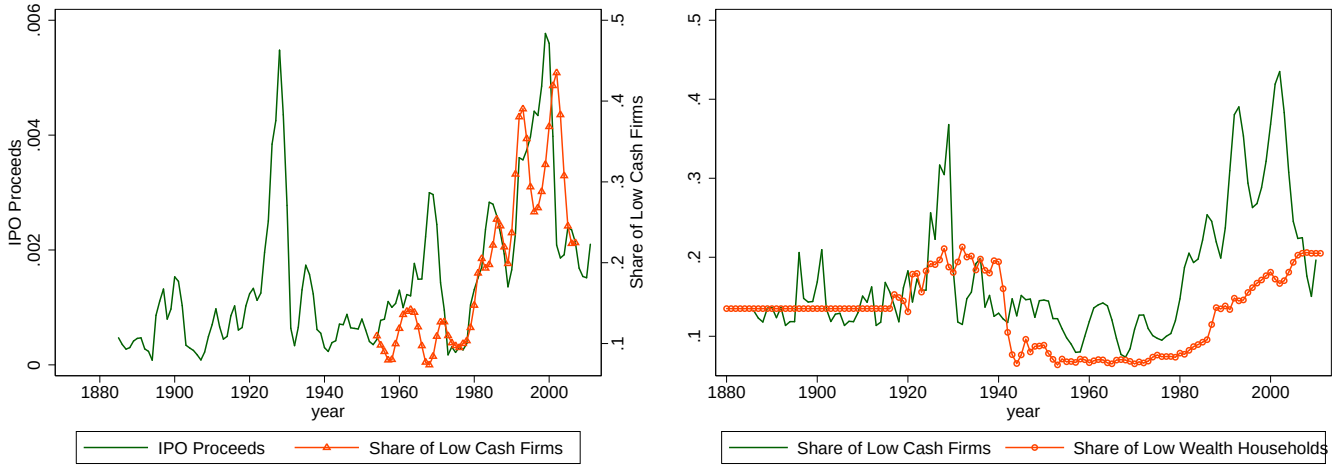
that they also require more intermediation services per-unit of borrowing. As a result, the intermediation intensity is an increasing function of the share of low cash firms. In the Appendix, I show that

$$\bar{\mu}_k(s_k) = \frac{\mu_k^h + (1+r)\left(x^h - x^l\right) s_k}{1 - x^h + \left(x^h - x^l\right) s_k}$$

where $s_k \equiv \frac{k_l}{k_l + k_h}$ is the fraction of high monitoring (low-cash) firms in aggregate investment. If $s_k = 0$, we have the benchmark model with $\bar{x} = x^h$ and $\bar{\mu}_k = \frac{\mu_k^h}{1-x^h}$ which is the monitoring intensity for the high type.[19]

The next step is to implement the adjustment. This requires a calibration of some parameters as well as empirical measure of s_k. Philippon (2008) uses Compustat to construct an empirical proxy for s_k, namely the share of aggregate investment that is done by firms that must borrow more than 3/4 of their capital spending. The measure is displayed in the left panel of Figure 10. Since it is based on Compustat, it is available only from 1950 onwards. Figure 10 also shows IPO proceeds, based on the work of Jovanovic and Rousseau (2001) and Ritter (2011). The two series are highly correlated in the post-war period, and I use the IPO series to extend the low cash share series before 1950. As argued by Jovanovic and Rousseau (2001), the IPO market of the 1920s was remarkably active, even compared to the one of the 1990s. IPO firms were of similar ages, and the proceeds (as share of GDP) were comparable. The next step is to estimate the adjustment. To do so I calibrate the complete model with household finance.

Figure 10: Inputs for Quality Adjustments



Notes: The share of low cash firms is from Philippon (2008). The IPO series is a three-year centered moving average of IPO proceeds over GDP. Sources are Jovanovic and Rousseau (2001) and Ritter (2011). The Share of Low Wealth Households in borrowing is estimated from Census Data and from Piketty and Saez (2003).

[19] In this case the high type would be the marginal type with a monitoring requirement per firm of $\mu_k^h = \frac{1+r}{1+\zeta}\left(\xi_k - x^h\right)$.

4.2 Household Finance and Calibration

It is also important to take into account heterogeneity among households. On a per-dollar basis, it is more expensive to lend to poor households than to wealthy ones, and we know that relatively poor households have gained access to credit in recent years. Using the Survey of Consumer Finances, Moore and Palumbo (2010) document that between 1989 and 2007 the fraction of households with positive debt balances increases from 72% to 77%. This increase is concentrated at the bottom of the income distribution. For households in the 0-40 percentiles of income, the fraction with some debt outstanding goes from 53% to 61% between 1989 and 2007. In the mortgage market, Mayer and Pence (2008) show that subprime originations account for 15% to 20% of all HMDA originations in 2005.

To capture the impact of household heterogeneity, I use a model similar to the one I have used above for firms. There are two types of households $i = l, h$. They differ by their labor incomes when they are young: $\eta_1^l < \eta_1^h$. The low-cash households end up using more intermediation per unit of borrowing than their richer counterparts, as explained in the appendix. In the model, the share of low cash households in the credit market depends on inequality among households, i.e. to $\eta_1^h - \eta_1^l$. Empirically, I use historical data from Piketty and Saez (2003) to measure inequality. Together with the data on low cash firms described above, this gives two time series that measure changes in the composition of borrowers over time. These two series are displayed on the right panel of Figure 10.

The last step is to calibrate the model and construct the required quality adjustments. Table 3 presents the parameters. Some parameters (in the first row) are set using values that are standard in the literature. The finance-specific parameters are chosen to match a set of moments. The moments are presented in the second row of the table, and the implied parameters in the last row. I calibrate the model using 1980 as a reference year. Low cash firms cover about 10% of their expenditures, so I set $x^l = 0.1$. They account for 20% of investment in 1980 (see Figure 10), so I use $s_k = 0.20$ as a target. I also target the ratios $\frac{B_k}{Y}; \frac{B_c}{Y}; \frac{M}{Y}$ in 1980, and the finance share of income, $\phi = 5\%$. For households one must keep in mind that, in the model, the long term households earn all the capital income, and the short term households earn only labor income. Within this group, I assume that η_1^l is the income of 30th percentile and η_1^h that of the 70th percentile. Using Census data (Jones and Weinberg, 2000), I calibrate the be ratio $\eta_1^h/\eta_1^l = 2.5$ for 1980.

Using the model, I can now adjust the output series of Figure 7. The adjustment factors are presented in the left panel of Figure 11. The normalization is such that when the low cash share is zero, the adjustment factor is one. When the low cash firms account for 20% of investment, the adjustment factor is around 1.18. This means that the correctly adjusted output is approximately 18% higher than the unadjusted one.

The bottom panel of Figure 11 shows the output measures with quality adjustments. As argued earlier the quality adjustments are small between 1947 and 1973, which makes it an ideal period to test the constant returns to scale assumption. Adjustments to consumer credit are small until 1995, while adjustments to business intermediation are important both before 1935 and after 1980. Given that the adjustment is roughly zero in 1960, this implies

Table 3: Calibration and Estimation

Direct Calibration

Rate	Deprec.	Growth	Labor Sh.	CRRA
$r = 0.05$	$\delta = 0.1$	$\gamma = 0.02$	$\alpha = 0.7$	$\rho = 1$

Targets for Estimation (as of 1980)

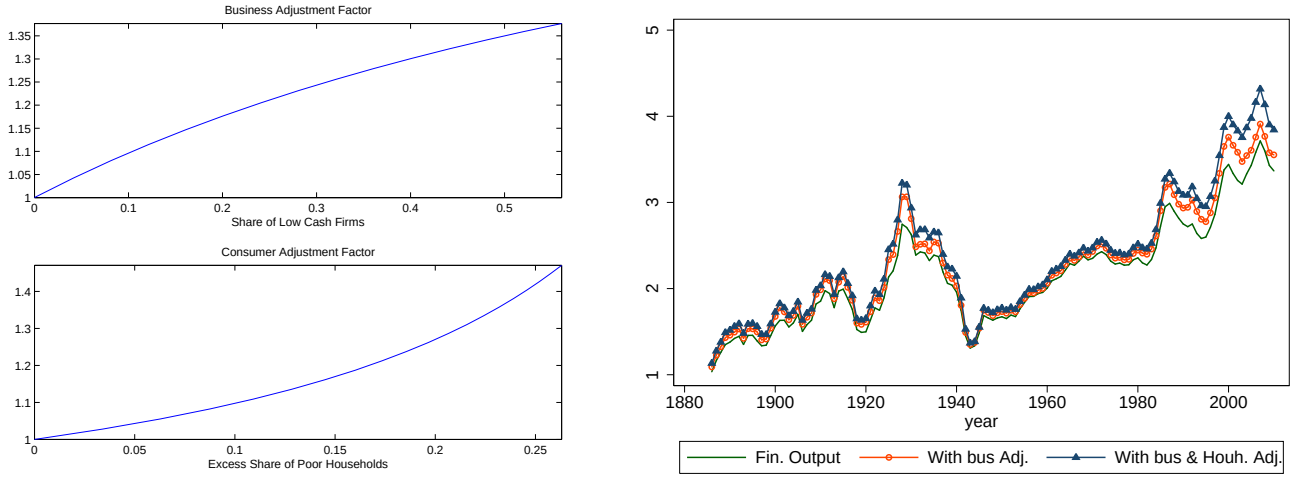
Business	Household	Liquidity	Fin. Share	l-Firms
$\frac{B_k}{Y} = 0.90$	$\frac{B_c}{Y} = 0.75$	$\frac{M}{Y} = 0.75$	$\phi = 0.05$	$s_k = 0.20$

Implied Parameters

$x^h = 0.56$	$\eta_2 = 0.73$	$\nu = 0.0125$	$\zeta = 0.21$	$\xi = 0.92$

that the unadjusted measure of business intermediation in Figure 7 underestimates the increase in intermediation between 1960 and 1980 by about 20%.

Figure 11: Quality Adjustments



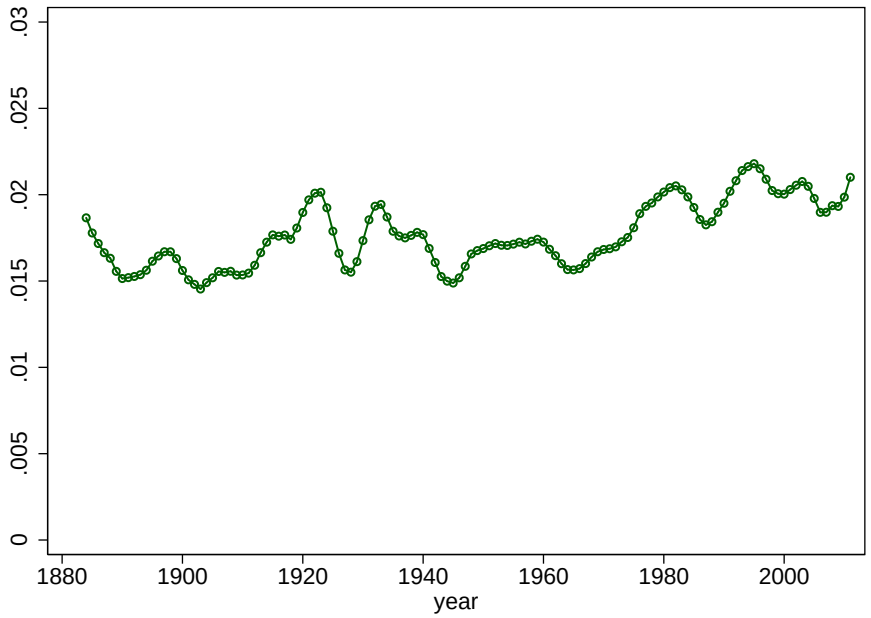
Notes: Adjustments are computed using the calibrated model (Table 3).

4.3 Adjusted Unit Cost

Figure 12 estimates the cost of financial intermediation, defined as income divided by output. For output, I use the composite measure of equation (8) adjusted for changes in the composition of borrowers, as in Figure 11. Figure 12 is the main contribution of the paper. It brings together the theory and the historical/empirical work of Section 3. There are two main points to take away. The first is that the ratio is remarkably stable. Recall that we start from an income series that fluctuates by a factor of 5 (9% relative to less than 2%). All the debt, deposit and equity series also vary a lot over time. But their ratio is stable. The series in Figure 12 has a mean of 1.76%, a standard deviation of 18.5 basis points, a minimum of 1.46% and a maximum of 2.18%. The series in Figure 8 has a mean of 1.935%, and a standard deviation of 24.3 basis points. The adjustments reduce the volatility of the unit cost measure by a quarter.

The second main point is that the finance cost index has increased since the mid-1970s. This is counter-intuitive. If anything, the development of information technologies (IT) over the past 40 years should have disproportionately increased efficiency in the finance industry. How is it possible for today's finance industry not to be significantly more efficient than the finance industry of John Pierpont Morgan a century ago? Figure 12 presents a puzzle for future research. In the next (and last) section, I discuss some recent research that sheds light on this puzzle.

Figure 12: Quality-Adjusted Unit Cost of Financial Intermediation



Notes: Total intermediation costs divided by quality-adjusted composite measure of financial intermediation output.

5 Discussion

5.1 Information Technology

Advances in information technology (IT) should lower the physical transaction costs of buying, pooling and holding financial assets. This seems especially relevant for equity markets. An important point of the empirical work is that I measure equity at market value. In equilibrium, if the cost of holding a diversified portfolio of stocks goes down, then the value of the portfolio should go up. Notice that my measure would attribute the entire increase in the price-earnings ratio to an improvement in financial intermediation.

In the model, advances in IT lower the unit cost of financial intermediation, but not necessarily the amount spent on intermediation. The main reason is that a decrease in ζ can give access to finance to borrowers who were previously priced out. When this extensive margin is elastic, a drop is ζ can lead to an increase in the finance income share. Conversely, for services that are used by (almost) all agents (e.g., deposits and checking accounts, but also more recently mutual funds) a decrease in ζ should lead to a decrease in the finance income share. An apt analogy

is with retail and wholesale trade. Indeed, in these sectors Philippon (2012b) shows that larger IT investment coincides with lower (nominal) GDP shares. In finance, however, exactly the opposite happens: IT investment and the income share are positively related. From this perspective, IT seems to deepen the puzzle instead of solving it. One could argue, however, that it is through its impact on secondary market trading that IT really matters, as discussed below.

5.2 Price Discovery and Risk Sharing

Using the GDP share of finance to measure the costs of financial intermediation is fairly straightforward. It ignores hidden costs of systemic risk, but it captures all fees and spreads. The output measure developed above, however, might not fully capture the production of information via prices, and the provision of insurance. Going back to the theory, it is important to ask the following question: If improvements in financial intermediation lead to more informative prices or better risk sharing, where would these improvements be seen in equilibrium?

Price Discovery

The simplest way to test the hypothesis that prices have become more informative is to directly test the signal-to-noise ratio of asset prices. Bai, Philippon, and Savov (2011) ask if current firm-level equity and bond prices predict future firm productivity and if this forecasting power has changed over time. They find that the forecasting power has been remarkably stable over the past 50 years (for comparable sets of firms). In other words, while bid-ask spreads have decreased, and while many have claimed that financial markets have become more liquid, this does not appear to translate into "better" prices. I am not aware of direct evidence regarding other asset classes. For commodity prices, some practitioners (e.g., Hadas 2011) seem to argue that prices have become less informative. See also Tang and Xiong (2011) for recent research on commodity prices.

Derivatives and Risk Management

The market for financial derivatives is extremely large. Since these contracts are in zero net supply, however, they do not enter *directly* into my calculations of output for the finance industry. The question is: Should they? The answer is essentially no, because the benefits of derivatives are already *indirectly* taken into account.

One thing is clear: it would make no sense to count derivatives at face value.[20] Rather, one should take the perspective of standard economic theory and recognize that derivatives can add real value in one of two ways: (i) risk sharing; (ii) price discovery. I have already discussed price discovery. Let me now discuss risk management, by financial firms, and by non financial firms. Risk management among banks lowers intermediation costs as banks lay off excess risk inventories when making markets. This type of risk sharing among financial intermediaries does

[20] For instance, consider the following example. Without derivatives, corporation A borrows from bank B and bank B retains the credit and duration risks on its books. With derivatives, bank B buys insurance against credit risk from fund C using a CDS. The sum of B and C holds exactly the same risk. Absent other frictions, the two models are equivalent.

not create any bias in my measurements. Improvements in risk sharing among banks would simply lead to lower borrowing costs and cheaper financing, and this would be correctly captured by the model.[21]

For the non-financial sector, I have already explained in Section 2 how the model captures improvements in risk management. In equilibrium, risk management translates into a lower used cost of capital, more issuances and more investment. Better risk management could also increase TFP if high productivity projects are risky. Better risk management would then translate into higher average firm value. But since I measure business capital at market value, I already capture these effects. Also remember that the liquidity measure accounts for cash and liquidity management. I conclude that the business risk management functions of derivatives are unlikely to bias my estimates.[22]

Risk Sharing and Consumption Smoothing

At the household level, better risk sharing should lead to less consumption risk. As with business risk management, the relevant question is whether my measure captures risk sharing among households. Let me start with an example where the measure works well. There is evidence of improved consumption smoothing in the housing market. Gerardi, Rosen, and Willen (2010) find that the purchase price of a household's home predicts its future income. The link is stronger after 1985, which coincides with important innovations in the mortgage market. The increase in the relationship is more pronounced for households more likely to be credit constrained. The model of Section 4 captures correctly these effects: an improvement in household finance leads to more borrowing and better consumption smoothing, especially for relatively poor households. More generally, consumption smoothing that entails the creation of credit flows does not create a bias in my estimation. Smoothing over the life cycle fits into this category. My measures are also correct if consumers use credit cards to smooth out income shocks.

It is nonetheless useful to look directly at risk sharing. Income inequality has increased dramatically in the U.S. over the past 30 years. If financial markets improve risk sharing, however, one would expect consumption inequality to increase less than income inequality. This is a controversial issue, but Aguiar and Bils (2011) find that consumption inequality closely tracks income inequality over the period 1980-2007. It seems difficult to argue that risk sharing among households has improved significantly over time. It is also difficult to point to a financial innovation in the past 30 years that would have directly improved risk sharing opportunities among households.

I would like to conclude this section with some caveats. First, my measures would not correctly capture *all*

[21] To see why let us go back to the simple example. Suppose there are frictions that rationalize why B and C should be separate entities, and why they gain from trading with each other (i.e., B has a comparative advantage at managing duration risk, and C at managing credit risk). Then the existence of CDS contracts can improve risk sharing among intermediaries, lower the risk premia, and lead to a decrease in the borrowing costs of A. Hence, with free entry, the total income going to intermediaries B+C would decrease. This could then increase the demand for borrowing, as explained earlier. All these effects would be captured by the model: either borrowing costs would go down, or borrowing volumes would go up, or both. In all cases, my approach would register an increase in efficiency.

[22] In any case, I am not aware of any direct evidence suggesting significant improvements in corporate risk management. One obvious index, that of precautionary savings by businesses, suggest even the opposite: corporate cash holdings have increased over the past 30 years. I am also not aware of direct evidence of credit derivatives leading to better risk management. For instance, it is commonly believed that hedging represents only a small fraction of all trades in the CDS market.

types of improvements in financial intermediation. For instance, the measures would be biased if households started trading large quantities of derivatives to improve risk sharing. There would be an improvement in welfare without an increase in credit flows. However, in practice, derivatives are mostly used by businesses, and for businesses the user cost framework correctly captures the value of risk management. And when households do use instruments for consumption smoothing, these instruments typically are not derivatives, and therefore are captured by the flow of funds.

A second caveat I have treated risk management and price discovery as two separate issues, but they need not be. In DeMarzo and Duffie (1991) for instance, financial hedging is fundamentally linked to private information about firm value, and in DeMarzo and Duffie (1995) hedging interacts with incentives and accounting disclosure. Those complex and fascinating issues are beyond the scope of this paper.

5.3 Trading

At this point, we are left with a puzzle. Even accounting for all the financial assets created in the U.S., the unit cost of intermediation appears to have increased. Finance has obviously benefited from the IT revolution and this has certainly lowered the cost of retail finance. So why is the non-financial sector still transferring so much income to the financial sector?

One proximate cause might be secondary market trading. Trading costs have decreased (Hasbrouck, 2009), but trading volumes have increased even more, and active fund management is expensive. French (2008) estimates that investors spend 0.67% of asset value trying (in vain on average, by definition) to beat the market. French's calculations are only for the equity market. In Figure 12, a drop in the intermediation cost index of 50 to 60 basis points would indeed bring it back towards its historical average. With output at 4 times GDP, this suggests that about 2% of GDP, or about $280 billions annually, are either wasted or at least difficult to account for.

Why do people trade so much? Financial economics does not appear to have a good explanation yet. An obvious but unsettling reason might be that they simply enjoy it. Another explanation is overconfidence, as in Odean (1998). Recent work by Glode, Green, and Lowery (2010) and Bolton, Santos, and Scheinkman (2011) explains why some type of informed trading might be excessive. Pagnotta and Philippon (2011) present a model where trading speed can be excessive. In these models, advances in IT do not necessarily improve the efficiency of financial markets.

6 Concluding Remarks

I have provided benchmark measures of production and efficiency for financial intermediation in the U.S. over the past 130 years. The cost of intermediation represents an annual spread of 2%. This spread is quite stable but it has increased over the past 30 years. This increase does not reflect better risk sharing or better information production. It represents a puzzle to the extent that one would expect information technology to lower the cost of intermediation

in finance, as it did in other sectors. The increasing cost might reflect inefficiencies driven by zero-sum trading activities or by inefficient regulations. More research is needed to answer these important questions.

Appendix

A Proof of Proposition 1

Let us start with the market clearing conditions. Long-lived households own the capital stock: $S_t = K_{t+1} + B_t^c$. Adding up the budget constraints we have

$$W_t + (1+r) S_{t-1} + (1-\psi_c) B_t^c - (1+r) B_{t-1}^c - \psi_m M_t = C_{0t} + C_{1t} + C_{2t} + S_t$$

The two sides of GDP are

$$\begin{aligned} Y_t &= W_t + (r + \delta + \psi_k) K_t \\ Y_t &= K_{t+1} + C_{0t} + C_{1t} + C_{2t} - (1 - \delta - \psi_k) K_t + \psi_m M_t + \psi_c B_t^c \end{aligned}$$

Combining them we get

$$K_{t+1} + C_{0t} + C_{1t} + C_{2t} = W_t + (1+r) K_t - \psi_c B_t^c - \psi_m M_t$$

Combining with the budget constraint and capital market equilibrium we get $(1-\psi_c) B_t^c = -\psi_c B_t^c + B_t^c$ which is simply the zero profit condition for consumer credit intermediaries. On the balanced growth path we have

$$\begin{aligned} y &= (1+\gamma) k + c_0 + c_1 + c_2 - (1 - \delta - \psi_k) k + \psi_m m + \psi_c b^c \\ y &= (\gamma + \delta + \psi_k) k + c_0 + c_1 + c_2 + \psi_m m + \psi_c b^c \end{aligned}$$

Aggregate consumption and money demand are

$$\begin{aligned} c_0 + c_1 + c_2 &= \frac{1}{1+\nu} (w - \psi_c b^c + (r - \gamma) k) \\ m &= \frac{\nu}{\psi_m} c \end{aligned}$$

The budget constraints of short-lived and long-lived households imply

$$\begin{aligned} (1+\nu)(c_1 + c_2) &= w + (1-\psi_c) b^c - \frac{1+r}{1+\gamma} b^c = w - \psi_c b^c - \frac{r-\gamma}{1+\gamma} b^c \\ (1+\nu) c_0 &= (r-\gamma) \left(k + \frac{b^c}{1+\gamma} \right) \end{aligned}$$

We have the Euler conditions

$$\begin{aligned} \beta (1+r) &= (1+\gamma)^\theta \\ c_1 &= (1-\psi_c)^{\frac{1}{\theta}} c_2 \end{aligned}$$

and, using $n = 1$,

$$\begin{aligned} k &= \left(\frac{1-\alpha}{r + \delta + (1-\bar{x}) \psi_k} \right)^{\frac{1}{\alpha}} \\ w &= \alpha k^{1-\alpha}. \end{aligned}$$

It is clear from these equations that the solution is unique. That homogeneity in production is required for balanced growth is not surprising. What is more interesting is that it is sufficient even if the production technologies differ between the financial and non-financial sectors (see Acemoglu and Guerrieri (2008) and Philippon (2012a) for detailed discussions). Regarding liquidity demand, balanced growth comes from the assumed preferences, as discussed in Lucas (2000). Under these assumptions the model predicts no income effect (i.e., no mechanical tendency for the finance income share to grow with per-capita GDP).

Proposition 1 contains 4 results of comparative statics:

(i) Suppose $\bar{\mu}_k$ goes down or $\bar{x}$ goes up. Then $(1-\bar{x})\,\psi_k$ goes down, and k, w, and k/y increase. We have

$$b_c = \frac{(1-\psi_c)^{\frac{1}{\theta}}\,\eta_2 - \eta_1}{1-\psi_c+(1-\psi_c)^{\frac{1}{\theta}}\,\frac{1+r}{1+\gamma}}w.$$

Since w/y is constant, so is b_c/y. But c/y increases because k/y increases, and so does m/y.

(ii) Suppose $\bar{\mu}_c$ goes down. Then ψ_c goes down and b_c, c and m go up. Since the user cost of capital is not affected, k and w are constant.

(iii) Essentially the same as (ii).

(iv) Suppose η_2 goes up. Then b_c goes up while k, w and y are unchanged. Then c and m go down, but one can check that $\bar{\mu}_c b_c + \bar{\mu}_m m$ goes up. We have

$$\begin{aligned} y^\phi &= \bar{\mu}_c b_c + \bar{\mu}_m m + \bar{\mu}_k b_k \\ \phi &= \frac{\zeta y^\phi}{y+\zeta y^\phi} \end{aligned}$$

Therefore y^ϕ and ϕ increase. Suppose now that ζ goes down. There are two effects: the direct effect and the fact that y^ϕ and y go up. So the impact but ϕ is ambiguous. See Philippon (2012a) for an analysis of functional forms and of the impact of heterogeneity.

B Quality Adjustments

This section presents the model with heterogenous firms and households.

B.1 Firms

There are k firms. Firm i is endowed with $x_i A$. Each firm needs to borrow $(1-x_i)A$ to operate a technology that produces according to $f(n) = An^\alpha$. With Cobb-Douglass technology, we get net income

$$\begin{aligned} \pi(w) &= (1-\alpha)\left(\frac{\alpha}{w}\right)^{\frac{\alpha}{1-\alpha}} \\ n &= \left(\frac{\alpha}{w}\right)^{\frac{1}{1-\alpha}} \end{aligned}$$

For any value of x, the payoffs are as follows. If the firm behaves well, it pays back its outside investors $(1+r)(1-x)$ and inside investors receive

$$\pi(w) + 1 - \delta - (1-x)(1+r) - \zeta\mu_k = \pi(w) - \delta - r - \zeta\mu_k + (1+r)\,x,$$

where ζ is the unit cost of intermediation and μ_k the quantity of monitoring used by the firm. Monitoring reduces the risk of diversion. If the firm cheats, outside investors receive nothing and inside investors keep $(1+r)\,\xi_k - (1+\zeta)\,\mu_k$, where ξ_k measures the degree of diversion.[23] The incentive constraint is therefore $\pi(w) - \delta - r - \zeta\mu_k + (1+r)\,x \geq (1+r)\,\xi_k - (1+\zeta)\,\mu_k$.

[23] this assumes that firms cannot divert bankers' fees. The analysis is essentially the same if they can, one must simply carry an extra term $\zeta\mu_k$ in the formulas.

The program of the firm is to minimize the cost of monitoring subject to the incentive and break-even constraints:

$$\begin{array}{rcl} \min\limits_{\mu_k \geq 0} \mu_k & & s.t. \\ \pi(w) - \delta - r + (1+r)\, x & \geq & (1+r)\, \xi_k - \mu_k, \\ \pi(w) & \geq & r + \delta + \zeta \mu_k. \end{array}$$

Firms always use the minimal amount of intermediation services:

$$\mu_k(x) = \max\left(0; r + \delta - \pi(w) + (1+r)\left(\xi_k - x\right)\right).$$

The second constraint binds for marginal firms, i.e., firms that are indifferent between entering and staying out.

We consider a model with two types of firms l and h, with $x^l < x^h$. We study equilibria where the number of high cash ventures is exogenously given by k_h and l is the marginal type. There is free entry of low types, therefore we have

$$\pi(w) = r + \delta + \zeta \mu_k^l,$$

and

$$\mu_k^l = \frac{1+r}{1+\zeta}\left(\xi_k - x^l\right).$$

This pins down the required profit rate, and therefore the equilibrium wage:

$$\pi(w) = r + \delta + \frac{\zeta}{1+\zeta}(1+r)\left(\xi_k - x^l\right).$$

The h-firms earns rents since $\pi > r + \delta + \zeta \mu_k^h$. The relative monitoring intensity is captured by

$$\mu_k^l - \mu_k^h = (1+r)\left(x^h - x^l\right)$$

Corporate finance intermediation services are measured by

$$y_k^\phi = \sum_j k_j \mu_k^j$$

To be consistent with our previous notations, we define the amount of monitoring services per unit of firm borrowing as

$$\bar{\mu}_k = \frac{y_k^\phi}{b_k}.$$

By definition, we have $b_k = k_h\left(1 - x^h\right) + k_l\left(1 - x^l\right)$, so

$$\bar{\mu}_k(s_k) = \frac{\mu_k^h + (1+r)\left(x^h - x^l\right) s_k}{1 - x^h + \left(x^h - x^l\right) s_k}$$

where $s_k \equiv \frac{k_l}{k_l + k_h}$ is the fraction of high monitoring (low-cash) firms in aggregate investment.

B.2 Households

There are two types of households $i = l, h$, with different labor incomes when young: $\eta_1^l < \eta_1^h$. Young households borrow but can run away with $\xi_c \eta_2 W_{t+1}$ if not monitored or with $\xi \eta_2 W_{t+1} - A_t \mu_c^i$ when $A_t \mu_c^i$ units of monitoring are used. The household's problem is

$$\max_{\mu_{c,t}^i, C_{1,t}^i, C_{2,t+1}^i, M_{1,t}, M_{2,t+1}, B_t^i} u\left(C_{1,t}, M_{1,t}\right) + \beta u\left(C_{2,t+1}, M_{2,t+1}\right),$$

subject to

$$\begin{aligned} C_{1,t} + \psi_{m,t} M_{1t} + \zeta A_t \mu^i_{c,t} &\leq \eta^i_1 W_t + B^i_t \\ C^i_{2,t+1} + \psi_{m,t+1} M_{2,t+1} &\leq \eta_2 W_{t+1} - (1 + r_{t+1}) B^i_t \\ (1 + r_{t+1}) B^i_t &\leq \eta_2 W_{t+1}(1 - \xi_c) + A_t \mu^i_{c,t} \end{aligned}$$

With the preferences assumed in the paper, this leads to the Euler equation (written with detrended variables):

$$c^i_1 = (1 - \zeta (1 + r))^{\frac{1}{\theta}} c^i_2$$

the budget constraints

$$\begin{aligned} (1 + \nu) c^i_1 &= \eta^i_1 w + b^i - \zeta \mu^i_c \\ (1 + \nu) c^i_2 &= \eta_2 w - \frac{1 + r}{1 + \gamma} b^i_c \end{aligned}$$

and the monitoring demand

$$\frac{1 + r}{1 + \gamma} b^i_c = \eta_2 w (1 - \xi_c) + \frac{\mu^i_c}{1 + \gamma}$$

We can then solve for μ^i_c and for b^i_c. Finally total monitoring per household's borrowing $\bar{\mu}_c$ is

$$\bar{\mu}_c (s_c) = \frac{s_c \mu^l_c + (1 - s_c) \mu^h_c}{s_c b^l_c + (1 - s_c) b^h_c}$$

where s_c is the share of poor households in the population.

Finally, for the calibration in the paper, I restrict the moral hazard parameters to be the same for businesses and consumers: $\xi_c = \xi_k$.

C JP Morgan 2010

According to its 2010 annual, total net revenue for JPM Co was $103 billion, 51b of interest income and 52b of non-interest income. The investment bank earned $26 billion, 15 from fixed income markets, 5 from equity markets, and a bit more than 6 in fees. Of the 26, non interest income accounted for 18, including 6.2b in fees (3.1 and 1.6 for debt and equity underwriting, and 1.5 for advisory fees), 8.4b from principal transactions, and 2.5b from asset management fees. For its private clients, the investment bank raised 440b in debt and 65b in equity. This suggests underwriting fees of $3.1/440 = 0.70\%$ for debt, and $1.6/65 = 2.46\%$ for equity. The cost of equity underwriting is therefore about 3.5 times the cost of debt underwriting. The bank also raised 90b for US governments and non-profits. The bank advised 311 announced M&A (a 16% market share). The bank also loaned or arranged 350b.

References

Acemoglu, D., and V. Guerrieri (2008): "Capital Deepening and Non-Balanced Economic Growth," *Journal of Political Economy*, 116(3), 467–498.

Acharya, V. V., L. H. Pedersen, T. Philippon, and M. Richardson (2009): "Measuring Systemic Risk," Working Paper NYU.

Adrian, T., and H. S. Shin (2008): "Financial Intermediary Leverage and Value at Risk," Federal Reserve Bank of New York Staff Reports, 338.

Aguiar, M., and M. Bils (2011): "Has Consumption Inequality Mirrored Income Inequality?," Working Paper, University of Rochester.

Almeida, H., and T. Philippon (2007): "The Risk-Adjusted Cost of Financial Distress," *Journal of Finance*, 62(6), 2557–2586.

Altinkilic, O., and R. S. Hansen (2000): "Are There Economies of Scale in Underwriting Fees? Evidence of Rising External Financing Costs," *Review of Financial Studies*, 13, 191–218.

Bai, J., T. Philippon, and A. Savov (2011): "Have Financial Markets Become More Informative?," .

Baker, M., and J. Wurgler (2000): "The Equity Share in New Issues and Aggregate Stock Returns," *Journal of Finance*.

Basu, S., R. Inklaar, and J. C. Wang (2011): "The Value of Risk: Measuring the Services of U.S. Commercial Banks," *Economic Inquiry*, 49(1), 226–245.

Baumol, W. J. (1967): "Macroeconomics of Unbalanced Growth: The Anatomy of the Urban Crisis," *American Economic Review*, 57, 415–426.

Beck, T., A. Demirguc-Kunt, and R. Levine (2011): *Financial Structure and Economic Growth: A Cross-Country Comparison of Banks Markets, and Development*chap. The Financial Structure Database, pp. 17–80. MIT Press, Cambridge.

Bekaert, G., C. R. Harvey, and R. Lumsdaine (2002): "Dating the Integration of World Capital Markets," *Journal of Financial Economics*, 65, 203–249.

Bernanke, B., M. Gertler, and S. Gilchrist (1999): "The financial accelerator in a quantitative business cycle framework," in *Handbook of Macroeconomics*, ed. by J. B. Taylor, and M. Woodford, vol. 1C. Elsevier Science, North Holland.

Bickenbach, F., E. Bode, D. Dohse, A. Hanley, and R. Schweickert (2009): "Adjustment After the Crisis: Will the Financial Sector Shrink?," Kiel Policy Brief.

Bolton, P., T. Santos, and J. Scheinkman (2011): "Cream Skimming in Financial Markets," Working Paper, Columbia University.

Brunnermeier, M., and L. Pedersen (2009): "Market Liquidity and Funding Liquidity," *Review of Financial Studies*, 22, 2201–2238.

Buera, F. J., J. P. Kaboski, and Y. Shin (2011): "Finance and Development: A Tale of Two Sectors," *The American Economic Review*, 101(5), 1964–2002.

Carter, S., S. Gartner, M. Haines, A. Olmstead, R. Sutch, and G. Wright (eds.) (2006): *Historical Statistics of the United States Millennial Edition Online*Cambridge University Press.

Chen, Q., I. Goldstein, and W. Jiang (2010): "Payoff complementarities and financial fragility: Evidence from mutual fund outflows," *Journal of Financial Economics*, 97(2), 239 – 262.

Christiano, L., and D. Ikeda (2011): "Government Policy, Credit Markets and Economic Activity," mimeo Northwestern.

Corsetti, G., K. Kuester, A. Meier, and G. J. Müller (2011): "Sovereign risk and the effects of fiscal retrenchment in deep recessions," .

Curdia, V., and M. Woodford (2009): "Conventional and Unconventional Monetary Policy," mimeo Columbia.

DeMarzo, P. M., and D. Duffie (1991): "Corporate financial hedging with proprietary information," *Journal of Economic Theory*, 53(2), 261–286.

——— (1995): "Corporate Incentives for Hedging and Hedge Accounting," *Review of Financial Studies*, 8(3), 743–71.

Diamond, D. W. (1984): "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, 51, 393–414.

Diamond, D. W., and P. H. Dybvig (1983): "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91, 401–419.

Diamond, D. W., and R. G. Rajan (2001): "Liquidity Risk, Liquidity Creation and Financial Fragility: A Theory of Banking," *Journal of Political Economy*, 109, 287–327.

Estevadeordal, A., B. Frantz, and A. M. Taylor (2003): "The Rise and Fall of World Trade, 1870-1939," *The Quarterly Journal of Economics*, 118(2), 359–407.

French, K. R. (2008): "Presidential Address: The Cost of Active Investing," *The Journal of Finance*, 63(4), 1537–1573.

Gennaioli, N., A. Shleifer, and R. Vishny (2011): "A Model of Shadow Banking," Working Paper, Harvard University.

Gerardi, K. S., H. Rosen, and P. Willen (2010): "The Impact of Deregulation and Financial Innovation on Consumers: The Case of the Mortgage Market," *Journal of Finance*, 65(1), 333–360.

Gertler, M., and P. Karadi (2011): "A Model of Unconventional Monetary Policy," *Journal of Monetary Economics*, 58, 17–34, Working Paper NYU.

Gertler, M., and N. Kiyotaki (2010): "Financial Intermediation and Credit Policy in Business Cycle Analysis," Working Paper, NYU.

Glode, V., R. C. Green, and R. Lowery (2010): "Financial Expertise as an Arms Race," forthcoming in Journal of Finance.

Gorton, G., S. Lewellen, and A. Metrick (2012): "The Safe-Asset Share," NBER WP.

Gorton, G., and G. Pennacchi (1990): "Financial Intermediaries and Liquidity Creation," *Journal of Finance*, 45(1), 49–72.

Gorton, G., and A. Winton (2003): "Financial Intermediation," in *Handbook of the Economics of Finance*, ed. by G. M. Constantinides, M. Harris, and R. Stulz, pp. 431–552, North Holland. Elsevier.

Gorton, G. B., and A. Metrick (2012): "Securitized Banking and the Run on Repo," *Journal of Financial Economics*.

Greenwood, J., J. M. Sanchez, and C. Wang (2010): "Financing Development; The Role of Information Costs," *The American Economic Review*, 100(4), 1875–1891.

——— (2012): "Quantifying the Impact of Financial Development on Economic Development," Working Paper, University of Pennsylvani.

Greenwood, R., S. Hanson, and J. Stein (2011): "A Comparative-Advantage Approach to Government Debt Maturity," Working Paper Harvard University.

Guiso, L., P. Sapienza, and L. Zingales (2004): "The Role of Social Capital in Financial Development," *American Economic Review*, 94, 526–556.

Hadas, E. (2011): "Commodity prices are failing New Zealand test," *Financial Times*.

Haldane, A., S. Brennan, and V. Madouros (2010): "What is the contribution of the financial sector: Miracle or mirage?," in *The Future of Finance: The LSE Report*, ed. by A. T. et al. LSE.

Hall, R. E. (2011): "The High Sensitivity of Economic Activity to Financial Frictions," *Economic Journal*.

Hasbrouck, J. (2009): "Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data," *Journal of Finance*, 64(3), 1445–1477.

He, Z., and A. Krishnamurthy (2012): "A Model of Capital and Crises," *Review of Economic Studies*.

Holmström, B., and J. Tirole (1997): "Financial Intermediation, Loanable Funds and the Real Sector," *Quarterly Journal of Economics*, 112, 663–691.

Jones, A., and D. Weinberg (2000): "The Changing Shape of the Nation's Income Distribution: 1947-1998," Discussion paper, US Census Bureau.

Jovanovic, B., and P. L. Rousseau (2001): "Why Wait? A Century of Life Before IPO," *AER paper and proceedings*, 91, 336–341.

Jovanovic, B., and P. L. Rousseau (2005): "General Purpose Technologies," in *Handbook of Economic Growth*, ed. by P. Aghion, and S. Durlauf. Elsevier, Chapter 18.

Kaplan, S. N., and J. Rauh (2010): "Wall Street and Main Street: What Contributes to the Rise in the Highest Incomes?," *Review of Financial Studies*, 23(3), 1004–1050.

Kashyap, A., R. Rajan, and J. Stein (2002): "Banks as Liquidity Providers: An Explanation for the Coexistence of Lending and Deposit-Taking," *Journal of Finance*, 57, 33–73.

Kiyotaki, N., and J. Moore (2008): "Liquidity, Business Cycles and Monetary Policy," mimeo, Princeton.

Krishnamurthy, A. (2009): "Amplification Mechanisms in Liquidity Crises," *NBER WP*.

Krishnamurthy, A., and A. Vissing-Jorgensen (2010): "The Aggregate Demand for Treasury Debt," Working Paper Kellogg School of Management.

Kuznets, S. (1941): "National Income and Its Composition, 1919-1938," Discussion paper, National Bureau of Economic Research.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. W. Vishny (1998): "Law and Finance," *Journal of Political Economy*, 106, 1113–1155.

Lucas, R. E. J. (2000): "Inflation and Welfare," *Econometrica*, 68(2), 247–274.

Lucas, R. E. J., and N. L. Stokey (1987): "Money and Interest in a Cash-in-Advance Economy," *Econometrica,*, 55(3), 491–513.

Martin, R. F. (1939): *National Income in the United States, 1799-1938*. National Industrial Conference Board,.

Mayer, C., and K. Pence (2008): "Subprime Mortgages: What, Where, and to Whom?," Staff Paper Federal Reserve Board.

Mehra, R., F. Piguillem, and E. C. Prescott (2011): "Costly financial intermediation in neoclassical growth theory," *Quantitative Economics*, 2(1), 1–36.

Mehra, R., and E. C. Prescott (1985): "The Equity Premium: a Puzzle," *Journal of Monetary Economics*, 15, 145–161.

Merton, R. C. (1995): "A Functional Perspective of Financial Intermediation," *Financial Management*, 24, 23–41.

Midrigan, V., and D. Y. Xu (2011): "Finance and Misallocation: Evidence from Plant-Level Data," Working Paper, NYU.

Moore, J. (2011): "Leverage Stacks and the Financial System," Ross Prize Lecture.

Moore, K. B., and M. G. Palumbo (2010): "The Finances of American Households in the Past Three Recessions: Evidence from the Survey of Consumer Finances," Staff Paper Federal Reserve Board.

Obstfeld, M., and A. M. Taylor (2002): "Globalization and Capital Markets," NBER WP 8846.

Odean, T. (1998): "Volume, Volatility, Price, and Profit when all traders are above average," *Journal of Finance*, 53, 1887–1934.

O'Mahony, M., and M. P. Timmer (2009): "Output, Input and Productivity Measures at the Industry Level: The EU KLEMS Database," *The Economic Journal*, 119(538), F374–F403.

Pagnotta, E., and T. Philippon (2011): "Competing on Speed," NBER WP 17652.

Philippon, T. (2008): "The Evolution of the US Financial Industry from 1860 to 2007: Theory and Evidence," NBER WP No. 13405.

——— (2012a): "Equilibrium Financial Intermediation," Working Paper NYU.

——— (2012b): "Finance vs. Wal-Mart: Why are Financial Services so Expensive?," in *Lessons from the Crisis*, ed. by A. Blinder, A. Lo, and R. Solow.

Philippon, T., and A. Reshef (2007): "Wages and Human Capital in the U.S. Financial Industry: 1909-2006," NBER WP 13437.

Piketty, T., and E. Saez (2003): "Income Inequality in the United States, 1913-1998," *Quarterly Journal of Economics*, 118(1), 1–39.

Pozsar, Z., T. Adrian, A. Ashcraft, and H. Boesky (2010): "Shadow Banking," NY Fed Staff Report.

Rajan, R. G., and L. Zingales (1998): "Financial Dependence and Growth",," *American Economic Review*, 88, 559–586.

Reinhart, C. M., and K. S. Rogoff (2009): *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press.

Rhodes-Kropf, M., and D. T. Robinson (2008): "The Market for Mergers and the Boundaries of the Firm," *The Journal of Finance*, 63(3), 1169–1211.

Ritter, J. R. (2011): "Initial Public Offerings," mimeo available online.

Sargent, T. J., and B. D. Smith (2009): "The Timing of Tax Collections and the Structure of "Irrelevance" Theorems in a Cash-in-Advance Model," mimeo NYU.

Scharfstein, D. (1988): "The Disciplinary Role of Takeovers," *The Review of Economic Studies*, 55, 185–199.

Scharfstein, D., and A. Sunderam (2011): "The Economics of Housing Finance Reform," Working Paper Harvard Business School.

Schularick, M., and A. M. Taylor (forthcoming): "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles and Financial Crises, 1870–2008," *American Economic Review*.

Sidrauski, M. (1967): "Rational Choice and Patterns of Growth in a Monetary Economy," *American Economic Review*, 57, 534–544.

Sirmans, C. F., and J. D. Benjamin (1990): "Pricing fixed rate mortgages: Some empirical evidence," *Journal of Financial Services Research*, 4, 191–202, 10.1007/BF00365422.

Stein, J. (2012): "Monetary Policy as Financial-Stability Regulation," *The Quarterly Journal of Economics*, 127(1), 57–95.

Tang, K., and W. Xiong (2011): "Index Investment and Financialization of Commodities," NBER WP 16385.

Wang, J. C., S. Basu, and J. G. Fernald (2008): "A General-Equilibrium Asset-Pricing Approach to the Measurement of Nominal and Real Bank Output," NBER WP No. 14616.