Finance and the Preservation of Wealth

Nicola Gennaioli, Andrei Shleifer, and Robert Vishny[1]

May 2013

Abstract

We introduce the model of asset management developed in Gennaioli, Shleifer, and Vishny (2012) into a Solow-style neoclassical growth model with diminishing returns to capital. Savers rely on trusted intermediaries to manage their wealth (claims on capital stock), who can charge fees above costs to trusting investors. In this model, the size of the financial sector rises with aggregate wealth, and wealth grows relative to GDP. As a consequence, the ratio of financial income to GDP rises over time, even though fees for given financial services decline. Because the size of the financial sector fluctuates with changes in investor trust, the model can account for the sharp decline of finance in the Great Depression, as well as its slow recovery afterwards. Entry by financial intermediaries as wealth increased in recent years may have further deepened investor trust and encouraged growth of financial income.

[1] Universita' Bocconi and CREI, Harvard University, and University of Chicago, respectively. We are grateful to Charles-Henri Weymuller for extremely helpful comments.

1.Introduction.

Philippon (2012) documents the astonishing rise of the share of GDP coming from the financial sector since World War II (Figure 1). Financial income rose from about 2% of the total in the 1940s to close to 8% at the time of the financial crisis. Philippon and Reshef (2013) document similar trends in many other developed countries. Greenwood and Scharfstein (2013) show further that, at least in the last 30 years, much of this rise of finance in the United States comes from financial services to consumers, especially asset management and credit intermediation of mortgages and consumer loans.

The rapid growth of the financial sector has proved difficult to explain. Perhaps productivity in finance, as in other services, does not grow as fast as that in other sectors, so we see a manifestation of the Baumol (1967) disease. However, finance has grown relative to other services (Philippon and Reshef 2013), and wages in finance have grown faster than in other service sectors (Philippon and Reshef 2012), inconsistent with this theory. Both Philippon (2012) and Greenwood and Scharfstein (2013) end up with skeptical interpretations of the growth of finance puzzle, focusing on socially unproductive innovation and rent-seeking.

We present a new model that addresses the evidence on the growth of finance. Ours is a dynamic Solow-style growth model with a financial sector serving the financial needs of individuals. The financial sector provides two services to savers. The first is wealth preservation: financial intermediaries enable investors to preserve their savings for future consumption. The second service is access to risky investment opportunities with superior expected return, which enables wealth to *grow* over time in expectation. As a byproduct of serving investors, the financial sector also provides investment resources to firms.

Our critical assumption is that investors need financial intermediaries to take advantage of these opportunities. On their own they only utilize highly inefficient self-storage, such as keeping

money in mattresses or in gold. Intermediaries offer savers knowledge of and access to financial products that they do not have otherwise. In Gennaioli, Shleifer, and Vishny (2012), we refer to the intermediaries providing such services, be they bankers, brokers, wealth planners, or money managers, as "money doctors." Our analogy captures the idea that even though generic investing in risky assets seems straightforward to economists and finance professors, it actually requires knowledge and confidence that most savers simply do not have. We follow Gennaioli et al in assuming that savers rely on money doctors to help them with risky financial choices. Moreover, as in that model and in Guiso, Sapienza, and Zingales (2004, 2008), trust in money doctors shapes investor risk-taking and competition in the industry. Specifically, we assume that investors are less anxious investing through intermediaries whom they know and trust. Intermediaries competitively set their fees to attract clients, but because some intermediaries have a "locational" advantage of being especially trusted by some clients, in equilibrium intermediaries charge positive fees that capture a share of expected returns on investments.

We analyze the dynamics of this model of capital accumulation and growth under the neoclassical assumption of diminishing returns to capital. We consider both the case of a fixed number of intermediaries, and of competitive entry by intermediaries, which yield similar results. The model predicts the growth of the finance share in GDP for the following reason. Recall that finance in our specification delivers two services to investors: growth and wealth preservation. With diminishing returns to capital, growth opportunities diminish over time, and so does the growth rate of GDP. Because it tracks returns to capital, this segment of financial services is a constant share of GDP. However, finance performs a second service, wealth preservation. Because the service of wealth preservation tracks wealth, and not GDP, the ratio of this part of financial income to GDP rises over time along with the ratio of wealth to GDP along the convergence path. Putting the two parts together, finance grows as a share of GDP, precisely because one of its main functions is to preserve the stock of wealth that, over time, is growing relative to GDP.

The basic neoclassical implication of the model, from which most others follow, is the growth of the ratio of wealth to GDP as the economy converges to the steady state. A recent presentation by Piketty and Zucman (2012) shows that over the long term this is indeed the case for many developed countries. We show for the US that the ratio of wealth to GDP has indeed increased for several plausible measures of wealth.

The model also delivers an important prediction that the cost of financial services falls over time. There are two reasons for this in the model, which focuses on investment fees. First, equilibrium fees are a share of expected returns in our model. As capital accumulates, and expected returns on capital decline, so do the expected returns on financial assets and management fees. In addition, in a free entry model, as the demand for financial services rises with wealth, there is entry by new money managers, who get "closer" to the investors, making finance more individualized. Competition between money managers also leads to the decline in fees over time.

Our model also sheds light on the empirical question of the evolution of costs of finance. Philippon (2012) suggests that unit costs of finance have remained constant. On the other hand, Greenwood and Scharfstein (2013) find that management fees, which are the costs of financial intermediation that our model specifically focuses on, have declined over time. Philippon and Reshef (2013) also point to the declining costs of financial intermediation in most countries outside the US. In our model, fees decline over time, in line with Greenwood-Scharfstein (2013) and Philippon-Reshef (2013). At the same time, the composition of investor portfolios tilts toward higher risk, intermediated (and therefore higher fee) financial products, which might explain why unit fees as measured by Philippon (2012) have not declined.

The gradual path of convergence to the steady state cannot account for two striking fluctuations in the size of the financial sector. First, Figure 1 illustrates the sharp decline in the financial sector in the Great Depression, which lasted for several decades. Second, the size of the financial sector since the Great Depression has grown relative to income and even wealth. To

account for such evidence, our model's central ingredient of trust as the lubricant of financial services proves critical. In our framework, the decline of finance in the Great Depression is explained as a decline in trust in the financial sector that took decades to rebuild. Indeed, it is difficult to imagine that technological parameters alone can account for the decline of finance share that took half a century to reverse, way longer than the recovery of productivity or of the real economy. Likewise, the growth of the financial sector since the Great Depression can be explained by growing investor trust in financial markets, as illustrated by the dramatic growth in stock market participation. In our model, trust may have increased both for exogenous reasons, as the memory of the Great Depression receded, and endogenous reasons, since increases in wealth encouraged entry by intermediaries, who got "closer" to their clients and therefore became more trusted.

Both Philippon (2012) and especially Greenwood and Scharfstein (2013) interpret their evidence as pointing to a rather skeptical view of the financial sector. Our model suggests otherwise, and the principal reason for that is the centrality of wealth preservation as the economy grows richer. Finance in our view is not so much an investment banking or trading service, as it is a service to savers, in the original Modigliani life cycle sense. The demand for that service of wealth preservation grows over time, and for that reason the growth of finance – even at rates faster than the growth of national income – is socially desirable in our model.

In Section 2, we describe our model. Section 3 presents the equilibrium in the financial sector. Section 4 considers the full equilibrium in the growth model, and discusses the relationship between the model's empirical implications and the available evidence. Section 5 extends the model to the case with endogenous entry of financial intermediaries. Section 6 concludes.

2. The Setup

This section introduces the key ingredients of the model: the household sector, financial intermediation, and the productive sector.

2.1 The Household Sector

The economy is inhabited by overlapping generations of young and old. Time starts at $t = 0$ and goes on forever. A generation born at time $t - 1$ contains a continuum of workers of size one, indexed by $i \in I_{t-1} \equiv [0,1]$. At $t - 1$, during their young age, these workers inelastically supply their unit labor endowments at the equilibrium wage w_{t-1}. The entire wage income is saved and invested as described below, and consumption takes place only in old age after investment income is received. At the end of t, the old generation dies without bequest. We describe an economy with no population growth or technological progress. These assumptions do not change our results, but an extension relaxing both assumptions in presented in Appendix B.1.

Workers can invest their resources in two ways. First, they can invest their income in safe self-storage. Each unit stored at $t - 1$ yields $\gamma \leq 1$ units at t, so that $1 - \gamma$ is lost in depreciation. We think of storage as an inefficient way to save on one's own, perhaps by holding cash or gold at home, or perhaps by keeping all the money in the bank. The case of $\gamma = 1$ captures a perfect self-storage technology. Second, a worker can hire a financial intermediary, whom we refer to as a money manager throughout, to invest his savings in a risky financial asset. At the beginning of time t, the money manager transforms a worker's resources (one for one) into capital, and rents it to firms, which use it to produce output at the end of time t. We later describe production in detail.

In the model, we draw a sharp distinction between self-storage, which requires no intermediation, and risky investments, which require money managers. In reality, the gradation is more continuous, from cash in mattresses and gold, to bank savings, to liquid market investments,

to illiquid investments such as private equity and hedge funds, with increasing amounts of intermediary attention (and cost). We view our sharp differentiation as a simplifying assumption.

There are a discrete number $m > 1$ of money managers in each generation, randomly selected from the young. A generic money manager active at time t is indexed by $j \in I_t$. This money manager charges his investors a profit-maximizing fee f_{jt} per unit of investment. At time t all managers invest in the same asset, which yields a stochastic gross return R_t with mean $\mathbb{E}\{R_t\}$ and variance σ_t, both of which are determined endogenously in equilibrium. A worker/saver born at time $t-1$ delegating at the beginning of time t his risky investment to manager j thus earns a net return $R_t - f_{jt}$. If the income share invested at time t in the risky asset is θ_t, the worker's consumption in old age is given by:

$$c_{it} = w_{t-1} \cdot \left[\gamma + \theta_t \cdot \left(R_t - \gamma - f_{jt}\right)\right].$$

Consumption increases in the excess return that risky financial assets earn relative to storage (net of the management fee). We impose the constraint $\theta_t \in (0,1)$ – which in Proposition 1 we verify to hold in equilibrium – because we are interested in cases where risk taking is interior.

After receiving the wage w_{t-1} at the end of period $t-1$, worker $i \in I_{t-1}$ chooses at the beginning of time t how much of that wage to invest in the risky asset, in storage, and which money manager $j \in \{1, \ldots, m\}$ to hire, so as to solve:

$$max_{j=1,\ldots,m,\theta_t \in (0,1)} \; w_{t-1} \cdot \left[\gamma + \theta_t \cdot \mathbb{E}\left(R_t - \gamma - f_{jt}\right) - a_{ij} \cdot \theta_t^2 \cdot \frac{\sigma_t}{2}\right]. \qquad (1)$$

The preferences of workers are mean-variance with respect to the *return* of their portfolio.[2] Critically, the utility of the investor i depends on the identity of manager j through the fee

[2]This objective function arises under quadratic utility when the agent's risk aversion is decreasing in his initial (pre-investment) wealth endowment, namely when:

$$u(c(W), W) = c(W) - \frac{b}{W} c(W)^2,$$

f_{jt} charged by j and through the manager-investor specific risk aversion parameter $a_{ij} > 1$, which we think of as the anxiety i experiences investing with j. As in Gennaioli et al. (2012), saver i sees risk as being more costly with manager j, anxiety a_{ij} as higher, the lower is the trust of i for j. Investors are less anxious when taking risk with more trusted managers, perhaps because they know them or their representatives personally, or perhaps because they are persuaded by advertisement. We thus capture lower trust of i in j by a higher value of the anxiety parameter a_{ij}.

2.2 Financial Intermediation

A worker's demand for the risky asset depends on his trust for different money managers and on the fees these managers charge. At each time t savers are uniformly distributed around the unit circle. Each manager j is also located along the circle at a constant distance $\Delta \equiv 1/m$ from the adjacent managers. The number of managers is exogenously fixed at m (we endogenize m in Section 5), and the trust of worker i in manager j is given by:

$$\frac{1}{a_{ij}} = \Gamma - d_{ij}, \tag{2}$$

where d_{ij} is the distance along the circle between the worker and the manager. The greater is the distance between worker i and manager j, the lower is trust and the higher is the worker's risk aversion.[3] Parameter $\Gamma \leq 1$ captures the maximal distance at which investor i is willing to delegate. If $d_{ij} > \Gamma$, the investor suffers infinite anxiety, namely $a_{ij} = \infty$, and so he only uses the storage

---

where consumption is the realized investment return, i.e. $c(W) = \check{R} \cdot W$. This utility function avoids the unappealing feature of standard quadratic utility that the share of wealth invested in the risky asset decreases with wealth W. It is also more tractable than constant relative risk aversion, which requires lognormal returns and analytical approximations that complicate optimal fee setting by money managers.

[3] Equation (2) captures investor trust in money managers. As a consequence, d_{ij} is zero when the money manager invests his own money, but not when a saver takes risk on his own. In fact, savers neither trust themselves nor other savers for risky investment. For simplicity, we assume that investors have zero trust (their risk aversion is infinite) with respect to homemade or non-professionally managed risk taking.

technology. Two managers located at distance Δ compete for some investors as long as $\Gamma > \Delta/2$. An investor located halfway between these two managers is willing to take some risk with either of them. When $\Gamma < \Delta/2$, investors located in the middle suffer infinite anxiety from hiring either manager. As a consequence, these investors do not take any risk and each manager has a small, captive, clientele. As we show below, whether general trust Γ is above or below $\Delta/2$ has interesting implications for the effect of competition on equilibrium fees.

At time t each money manager sets his fee for the generation of savers born at $t-1$. This results in a profile $\boldsymbol{f_t} \equiv (f_{1,t}, \dots, f_{m,t})$ of money managers' fees. Given this profile, each worker i chooses, based on his trust as described by (2), which manager to invest with and how much risky investment to undertake. The optimal policy of a worker $i \in I_{t-1}$ is summarized by a vector $\left[\theta_{ij}^*(\boldsymbol{f_t})\right]_{j=1,\dots,m}$ that takes nonzero value only for the manager to whom the worker delegates his risky investment. This vector is the solution of the investor's problem described in Equation (1). The optimal investment policy depends on time only through the fees $\boldsymbol{f_t}$ set by managers at time t. This implies that at a fee profile $\boldsymbol{f_t}$, the profit earned by a generic money manager j from time t investment is given by:

$$\pi_{jt}(\boldsymbol{f_t}) = f_{jt} \cdot \left[\int_i \theta_{i,j}^*(\boldsymbol{f_t})\, di\right] \cdot w_{t-1}. \tag{3}$$

We consider symmetric Nash equilibria in which each manager j sets the same optimal fee f_t^* identified by the condition:

$$f_t^* = argmax_{f_{jt}}\, \pi_{jt}(f_{jt}, f_{-jt} | f_{-jt} = f_t^*).$$

Before computing workers' investment decisions and management fees, we describe the production structure of the model.

2.3 The Productive Sector

There are two inputs: labor and capital, available in aggregate supply $\mathrm{L}_t = 1$ and K_t, respectively. We assume that capital can be converted back into consumption at no cost, but Appendix B.2 shows that our main results continue to hold when we relax this assumption. Inputs at time t are owned by workers (labor is owned by the young born at time t, capital is owned by the old who are born at time $t-1$) and hired by firms in competitive markets. The production technology is risky. If an individual firm hires k_t units of capital and l_t units of labor it produces:

$$F(k_t, l_t) = \varepsilon_t[\, k_t + A \cdot k_t^{\alpha} l_t^{1-\alpha}]. \tag{4}$$

In (4), ε_t is an i.i.d shock with mean $\mathbb{E}\{\varepsilon_t\} = 1$ and variance σ. Uncertainty is realized at the end of period t when output is produced. The value of a firm consists of two components. The first is its value added $\varepsilon_t \cdot A \cdot k_t^{\alpha} l_t^{1-\alpha}$, where A captures the firm's total factor productivity. The second component is the capital stock k_t used in production, which the firm returns to investors undepreciated (up to the stochastic shock ε_t).

At time t, before the shock ε_t is realized, firms hire capital and labor. Workers are hired on the spot market and are remunerated with a deterministic equilibrium wage w_t. The remuneration of capital is risky since it fully adjusts to the realization of the shock ε_t, and is paid to the holders of the firm's financial claims. These claims are bought by savers via money managers and pay an equilibrium return R_t with expected value $\mathbb{E}\{R_t\}$ and risk σ_t. The return R_t is thus determined by the firms' investment and by ε_t.

3. Equilibrium in the Money Management Sector

To solve a worker's portfolio problem and a manager's profit maximization problem, we take wages and expected asset returns as given. These variables are computed in the next section.

At time t, each saver – after collecting his period $t-1$ wages – optimally chooses a money manager and an amount of risky investment to solve Equation (1). If worker i selects money manager j, he invests in the risky asset a share $\theta_{ij}(f_{jt})$ of his wealth w_{t-1}. This share is given by:

$$\theta_{ij}(f_{jt}) = \frac{\mathbb{E}\left(R_t - \gamma - f_{jt}\right)}{a_{ij}\sigma_t}, \tag{5}$$

where $\theta_{ij}(f_{jt})$ is assumed to be in (0,1) (Proposition 1 verifies that this is the case). The saver invests $\theta_{ij}(f_{jt}) \cdot w_{t-1}$ in the risky asset and $\left[1 - \theta_{ij}(f_{jt})\right] \cdot w_{t-1}$ in storage. Risk taking increases in the excess return paid by the risky asset and in investor trust, but decreases in the risk σ_t of the financial asset. Consider now a worker's decision of which money manager to hire.

Figure 2 depicts the case with three managers, in which an investor i^* is located between managers j_1 and j_2. Consider the case when investors do not suffer infinite anxiety with either of the two closest managers, i.e., $\Gamma > \Delta/2$.

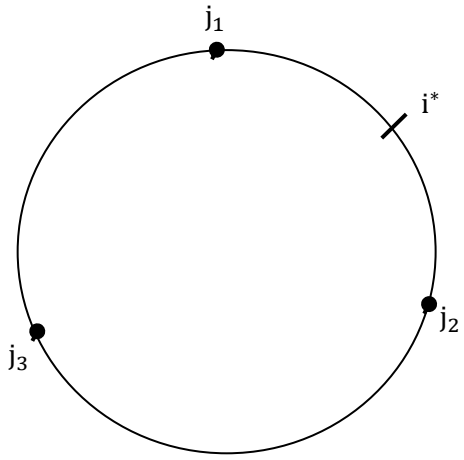


Figure 2

In this situation (and focusing on small deviations from a symmetric equilibrium), the investor chooses between the two closest managers j_1 and j_2. This implies that in setting his fee a generic manager, say j_2, competes for investors on his right against j_1 and for investors on his left against j_3. To see the implications of this logic for fee setting, consider the general case in which

an investor i chooses between his two closest managers j and j'. Denote the distance between investor i and his left-adjacent manager j by δ. Since the distance between managers is Δ, the distance between the same investor and his right-adjacent manager j' is $\Delta - \delta$. In light of Equation (2), these distances pin down in Equation (5) the investor's risky investment with either manager. By plugging these optimal risky investments into the investor's objective function of Equation (1), we can show that investor i obtains a higher utility by delegating his investment to manager j rather than to manager j' if and only if:

$$\delta \leq \delta\left(f_{jt}, f_{j't}\right) \equiv \Gamma - (2\Gamma - \Delta) \cdot \frac{1}{\left[\frac{\mathbb{E}\left(R_t - \gamma - f_{jt}\right)}{\mathbb{E}\left(R_t - \gamma - f_{j't}\right)}\right]^2 + 1}. \tag{6}$$

Investor i thus hires manager j when the above condition holds and manager j' otherwise. Intuitively, the investor delegates his risky portfolio to manager j when his trust in j is sufficiently high, as captured by a sufficiently small distance δ from j. Other things equal, delegation to manager j is also more likely when j charges a lower fee (f_{jt} is lower) and the competing manager j' charges a higher fee ($f_{j't}$ is higher).

Consider now optimal fee setting by manager j. With the assumed circular structure, a generic manager j competes for investors against his neighbors on the left and the right. Manager j attracts investors who – according to (6) – are sufficiently close to him. This implies that, if two competing managers j' and j'' set the equilibrium fees $f_{j't} = f_{j''t} = f_t^*$, then the profit of manager j from setting fee f_{jt} is given by:

$$2 \cdot w_{t-1} \cdot f_{jt} \cdot \int_0^{\delta\left(f_{jt}, f_t^*\right)} (\Gamma - \delta) \cdot \frac{\mathbb{E}\left(R_t - \gamma - f_{jt}\right)}{\sigma_t} \cdot d\delta,$$

where $\delta(f_{jt}, f_t^*)$ is the maximal distance at which an investor i prefers to hire manager j at fee f_{jt} to hiring his closest competitor at the equilibrium fee f_t^*. Maximization of the above profit function yields the (sufficient) first order condition:

$$\mathbb{E}(R_t - \gamma - 2f_{jt}) \cdot \int_0^{\delta(f_{jt}, f_t^*)} (\Gamma - \delta) \cdot d\delta + \frac{\partial \delta(f_{jt}, f_t^*)}{\partial f_{jt}} [\Gamma - \delta(f_{jt}, f_t^*)] \cdot f_{jt} \cdot \mathbb{E}(R_t - \gamma - f_{jt}) = 0.$$

At a symmetric equilibrium $f_{jt} = f_t^*$, we obtain the following result. All proofs appear in Appendix A.

Lemma 1 *The equilibrium fee at time t is given by:*

$$f_t^* = \left[\frac{\Delta}{\Gamma} - \left(\frac{\Delta}{2\Gamma}\right)^2\right] \cdot \frac{\mathbb{E}(R_t - \gamma)}{2} \equiv \varphi \cdot \mathbb{E}(R_t - \gamma). \quad (7)$$

where $\varphi < 1$. Management fees increase with the expected return on the risky asset. Furthermore, for $\Gamma > \Delta/2$ – which is equivalent to $m \geq 1/2\Gamma$, fees decrease in the number of managers m and in the generalized trust Γ that investors have in the financial sector as a whole.

From the empirical standpoint, unit fees in our model correspond to the ratio between aggregate financial sector income $f_t^* K_t$ and intermediated wealth K_t. As in Gennaioli et al. (2012), equilibrium fees capture a constant fraction of the excess return expected on the risky asset. This sharing rule is intuitive: managers extract part of the surplus they enable their trusting investors to access. The fraction φ of return extracted by managers decreases as trust in all managers Γ rises. When investors trust all managers, competition among them is very intense, which drives down fees. If $m \geq 1/2\Gamma$, fees also drop as the number of managers m rises. Intuitively, competition

between highly trusted managers lowers their market power and fees. Fees fall to zero as managers fill the entire circle, namely as $m \to \infty$. In the remainder, we focus on the case where $m \geq 1/2\Gamma$.[4]

By plugging Equation (7) into the optimal portfolio of Equation (5), we can show that investor i places in the risky asset a share of wealth given by:

$$\theta_{ij}(f_{jt}) = \frac{(1-\varphi)\cdot \mathbb{E}(R_t - \gamma)}{a_{ij}\sigma_t}.$$

In equilibrium, each investor hires the closest manager and each manager attracts the same amount of wealth. As a consequence, the aggregate share of wealth invested in the risky asset at t, which we denote by θ_t, is the product of the number of managers m and the share of wealth managed by each of them. This aggregate share is given by:

$$\theta_t \equiv \iint_{i,j} \theta_{ij}(f_{jt})\, didj = m \cdot 2 \cdot \left[(1-\varphi)\cdot \frac{\mathbb{E}(R_t - \gamma)}{\sigma_t} \cdot \int_0^{\frac{\Delta}{2}} (\Gamma - \delta) d\delta\right] =$$

$$= \frac{(1-\varphi)\mathbb{E}(R_t - \gamma)}{\sigma_t} \cdot \left(\Gamma - \frac{\Delta}{4}\right), \tag{8}$$

where the expression in square brackets captures the wealth share invested by the clients to the right of a manager. With symmetry, the wealth share managed by an individual manager is twice the amount in square brackets. Equation (8) says that the share of wealth invested in the risky asset increases in the asset's excess return (net of fees) per unit of risk, in overall trust Γ, and in the number of managers $1/\Delta$. As trust in money managers increases, fees drop, investors become less anxious and are willing to take more risk.

[4] The case $m < 1/2\Gamma$ has some interesting properties. When there are very few managers, a potentially large measure of investors located between any two managers does not take any risk. In this case, an entering manager could exploit monopoly (or quasi) monopoly profits by locating close to such excluded investors. In this scenario, entry of new money managers increases participation into risk taking while exerting limited (or no) downward pressure on the fees charged by existing managers.

4. General Equilibrium Dynamics

4.1 Production, Wages and Asset Returns

At time t, before observing ε_t, a generic firm hires labor and capital to maximize expected profits:

$$max_{k_t,l_t}\mathbb{E}\{\varepsilon_t \cdot k_t + \varepsilon_t \cdot A \cdot k_t^{\alpha} l_t^{1-\alpha} - w_t l_t - R_t k_t\},$$

which are equal to total output (inclusive of both value added and the capital stock) minus factor payments. Profit maximization yields the optimality conditions:

$$(1-\alpha)k_t^{\alpha} l_t^{-\alpha} = w_t,$$

$$1 + \alpha A k_t^{\alpha-1} l_t^{1-\alpha} = \mathbb{E}\{R_t\}.$$

The marginal product of labor is equated to the wage rate, and the average marginal product of capital is equated to the average (gross) return of financial assets $\mathbb{E}\{R_t\}$.

Because the real wage is deterministic, the firm's wage bill is also deterministic, given by $w_t l_t = (1-\alpha)Ak_t^{\alpha} l_t^{1-\alpha}$. The production function then implies that, upon the realization of a shock ε_t, the resources available to the firm's capital suppliers are $\varepsilon_t \cdot k_t + \varepsilon_t \cdot A \cdot k_t^{\alpha} l_t^{1-\alpha} - (1-\alpha)Ak_t^{\alpha} l_t^{1-\alpha}$. The rate of return per unit of capital in state ε_t is therefore given by:

$$R_t = \varepsilon_t + \varepsilon_t \cdot A \cdot k_t^{\alpha-1} l_t^{1-\alpha} - (1-\alpha)Ak_t^{\alpha-1} l_t^{1-\alpha}.$$

By taking the expected value of the above expression, one can immediately see that the expected return $\mathbb{E}\{R_t\}$ is equal to the average marginal product of capital $[1 + \alpha A k_t^{\alpha-1} l_t^{1-\alpha}]$, as in the first order condition above. With constant returns to scale, remunerating capital with the residual of output after the wage bill is paid is consistent with optimality. Evaluated at the aggregate endowments K_t and $L_t = 1$, the equilibrium wage and expected return are then given by:

$$(1-\alpha)AK_t^{\alpha} = w_t, \tag{9}$$

$$1 + \alpha AK_t^{\alpha-1} = \mathbb{E}\{R_t\}. \tag{10}$$

Furthermore, by using the above expression for R_t we can show that the variance of returns is equal to $\sigma_t = var(R_t) = \sigma[1 + AK_t^{\alpha-1}]^2$.

4.2 Evolution of the Financial Sector

We can now characterize the evolution of the economy. The total amount of risky investment at time t, which buys the aggregate capital stock K_t, is equal to the past aggregate wage bill w_{t-1}, times the share of this wealth invested with money managers:

$$K_t = \theta_t \cdot w_{t-1}.$$

Using Equations (9), we can rewrite this equation as:

$$K_t = \theta_t \cdot (1-\alpha)AK_{t-1}^{\alpha}. \tag{11}$$

By plugging equilibrium returns and variance into equation (8), we can compute the aggregate share of wealth invested in the risky asset, which is given by:

$$\theta_t = \frac{(1-\varphi)(1 + \alpha AK_t^{\alpha-1} - \gamma)}{\sigma[1 + AK_t^{\alpha-1}]^2} \cdot \left(\Gamma - \frac{\Delta}{4}\right). \tag{12}$$

Equations (11) and (12) fully characterize the dynamics of the economy. The law of motion of the capital stock in (11) is very similar to that obtained in a standard Solow model, with the main difference that now the amount of resources invested in the economy depends, through θ_t, on the equilibrium fees set by money managers and on the risk-return profile entailed by real investment.

In Appendix A we prove that, by combining (11) and (12) we obtain the following result:

Proposition 1 *If* $2\alpha > (1-\gamma)$, *there are two thresholds* $\bar{\sigma}$ *and* $\underline{\sigma}$, *with* $\bar{\sigma} > \underline{\sigma}$, *such that, for* $\sigma \in (\underline{\sigma}, \bar{\sigma})$ *the economy admits a unique nonzero steady state level of capital* K^* *at which individual risk taking is interior and aggregate risk taking is given by* $\theta^* < 1$. *The steady state is locally stable and displays the following properties:*

i) *The steady state capital stock weakly increases with the level of productivity and with the number of money managers, formally* $\partial K^*/\partial A > 0$, $\partial K^*/\partial m > 0$;

ii) *Risk taking increases with the level of productivity and with the number of money managers, formally* $\partial \theta^*/\partial A > 0$, $\partial \theta^*/\partial m > 0$.

When the volatility σ of the productivity shock is intermediate, the economy monotonically converges to a unique steady state level of financial intermediation and investment.[5] The steady state level of capital increases in productivity A. When investment becomes more productive, the wage earned by the young and the average return promised by money managers rise. Both effects increase financial intermediation, investment and output in the economy. An increase in the number $m = 1/\Delta$ of money managers also increases financial intermediation, investment and output in the steady state. There are two reasons for this. First, when m increases, investors can find a more trusted money manager, increasing – for given fees – their propensity to invest. Second, a higher m increases competition among money managers, reducing equilibrium fees and increasing for a given level of an investor's trust the investor's risk appetite.

[5] The role of production risk is intuitive. If σ is too low, people are very eager to invest in the risky asset. Some or all of them give all of their wealth to money managers, setting $\theta_{t-1}^{ij} = 1$. Condition $\sigma > \underline{\sigma}$ rules out this possibility. If σ is very high, the variance of the risky asset decreases very fast with the capital stock. This can be a source of multiplicity: some equilibria are characterized by low investment and high risk (which vindicates low investment), while other equilibria feature high investment and low risk (vindicating high investment). Condition $\sigma < \bar{\sigma}$ rules out this possibility.

One important property of our model is that the steady state is locally stable. That is, an economy starting below or above the steady state monotonically converges to it. Figure 3 graphically represents this convergence process.

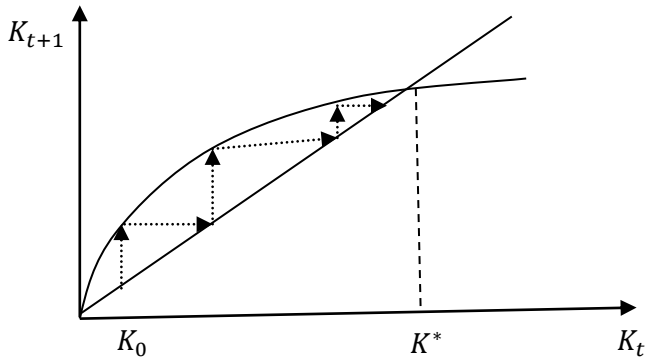


Figure 3

Similarly to the standard neoclassical growth models, the main source of stability is diminishing returns to capital. As the capital stock increases, wages and national income rise. This raises the demand for financial assets by savers. The increase in financial assets further increases the capital stock and thus output next period. The growth rate of the capital stock however declines over time, because new resources are invested at progressively lower returns. Growth stops eventually and the steady state is attained.[6]

This convergence process has interesting implications for the financial sector. In particular, how do fees and money management profits change as the economy grows over time? We address these issues below.

Corollary 1 *Suppose that the economy starts below the steady state, namely* $K_0 < K^*$. *During the transition to the steady state:*

[6] Unlike in the standard Solow model, however, diminishing returns are not enough to guarantee stability, because in our model risk taking by households increases as the capital stock grows. This phenomenon creates the possibility of explosive paths on which capital accumulation begets further risk taking and capital accumulation. The upper bound on the variance of shocks ensures stability by reducing the sensitivity of risk taking to the capital stock.

i) The unit fee charged by money managers, which is given by:

$$f_t^* = \varphi \cdot \mathbb{E}(R_t - \gamma) = \varphi \cdot [1 - \gamma + \alpha \cdot A \cdot K_t^{\alpha-1}], \quad (13)$$

decreases over time as capital accumulates.

ii) The total income of the financial sector increases over time, at a higher speed than value added. The ratio of financial sector income over value added (GDP), is given by:

$$\frac{\varphi \cdot \mathbb{E}(R_t - \gamma) \cdot K_t}{AK_t^{\alpha}} = \varphi \cdot \left[\left(\frac{1-\gamma}{A}\right) \cdot K_t^{1-\alpha} + \alpha\right]. \quad (14)$$

As the economy accumulates capital, there are more resources for financial intermediation. At the same time, diminishing returns to physical capital ($\alpha < 1$) imply that *ceteris paribus* these additional resources are employed at a lower marginal return. This explains why unit management fees fall along the transition. As capital deepening reduces the expected excess return on the risky asset, it also reduces the surplus that money managers can extract from investors. Declines in expected returns in turn reduce the rents available for money managers, as reflected in their unit fee.

Despite this reduction in unit fees, the aggregate income earned by money managers grows over time. This is because the growth in the size of the intermediated wealth K_t more than compensates for the reduction in unit fees, and causes financial sector income to rise over time. In our model financial sector income grows faster than value added, so the ratio of financial sector income to GDP grows over time.

To understand this result, recall that in our model financial sector income can be viewed as remuneration for two services. The first is a "wealth preservation" service: money managers allow savers to access investment opportunities which on average return the initial un-depreciated capital and are thus better than self-storage. The second is a "growth" service: money managers enable savers to earn part of the capital income generated by these productive investment opportunities. In equilibrium, money managers are remunerated for both services. The remuneration for wealth

preservation is equal to $\varphi(1-\gamma)K_t$, which is the product of the per unit of return fee φ times the surplus created by managers relative to riskless storage. Intuitively, wealth preservation is more expensive the worse is the return on riskless storage (i.e., the lower is γ). The remuneration for the growth service is equal to the per unit of return fee times capital income, namely $\varphi \cdot \alpha \cdot AK_t^{\alpha}$. This remuneration increases in total value added AK_t^{α} and in the share α of the value added that remunerates capital. As capital stock grows, the remuneration for both wealth preservation and growth services rises, in turn increasing the aggregate income of the financial sector.

Why does the total financial income grow faster than GDP? Consider the financial sector's growth services and wealth preservation separately. As a product of real growth opportunities, income from growth services grows at the same rate as GDP. Indeed, as shown by the second term in Equation (14), the remuneration for growth services is a constant fraction $\varphi \cdot \alpha$ of aggregate GDP. On the other hand, the first term in Equation (14) shows that the wealth preservation service grows faster than GDP. The remuneration for this service grows linearly with the capital stock K_t, which in turn grows faster than GDP: while value added is subject to decreasing returns, capital preservation is not. The fact that a portion of the financial services is dedicated to preserving the wealth of the economy, and not to the shrinking pool of new profitable investment projects,[7] causes the ratio of financial to total income to rise over time.

4.3. Empirical Predictions.

Our model yields several empirical predictions, some consistent with the available evidence, some new. To begin, our model turns the standard neoclassical prediction that the capital

[7] In Equation (14) we exclude storage services or capital preservation from the definition of GDP in the denominator order to illustrate most starkly our results. The growing relative size of the financial sector is however robust to alternative definitions. For instance, if one includes financial sector income into GDP, defining the latter as $\varphi(1-\gamma) \cdot K_t + AK_t^{\alpha}$, finance still increases as a share of GDP provided $\alpha\varphi < 1$, which always holds. But even if one takes the broadest notion of GDP to include the entire capital stock, namely $K_t + AK_t^{\alpha}$, finance would still increase as a share of GDP provided $\alpha < 1-\gamma$.

to income ratio increases during transitional growth into a novel rationale for Philippon's (2012) and Philippon and Reshef's (2013) finding that the financial sector grows relative to GDP. As past economic growth causes the accumulation of considerable wealth, workers (who own the financial claims on the capital stock) seek wealth preservation opportunities more attractive than self-storage. Trusted money managers provide access to such opportunities, which allows them to extract a fraction of the growing wealth. As the availability of profitable investment opportunities declines, the economy-wide wealth to income ratio rises, causing the *relative* income of the financial sector to rise over time. It is not that finance slows down economic growth by driving resources away from productive opportunities. It is rather the fact that productive opportunities diminish and past wealth must be preserved that allows the financial sector to expand relative to the productive sector.

The principal mechanism of the model, and thus its basic prediction, is the growing wealth to GDP ratio in the economy. Figure 4 presents this ratio, computed for both total and financial wealth, for the United States, and shows that it rises over time. A presentation by Piketty and Zucman (2012) shows for several developed countries that the ratio of wealth to GDP indeed grows over long stretches of time, although they do not connect this finding to the growth of finance.

The model so far also predicts the decline in management fees over time. As capital accumulates, the expected return on capital declines and equilibrium fees, which are a share of expected return, decline as well. It is precisely the same force of declining marginal product of capital that drives both the decline in fees and the increase in finance share. Greenwood and Scharfstein (2013) document the decline in management fees over time. However, Philippon (2012) finds that, in the United States, unit costs of finance have not declined. We attempt to reconcile this evidence after extending our model to incorporate entry by money managers in Section 5.

4.4. Fluctuations in the Size of the Financial Sector

We have so far focused on long term trends and have ignored fluctuations in the size of the financial sector, evident in Figure 1. Our model also allows us to analyze the short and long run responses of the financial sector to shocks. We compare the effects of two permanent shocks: a permanent drop in productivity A and a drop in the overall level of trust in the financial sector Γ, owing for instance to the erosion of investor confidence during a large scale financial crisis. Our model describes how the financial sector adjusts to these shocks.

Corollary 2 *Suppose that an economy is originally in a steady state* $K^*(\Gamma, A)$.

i) *Productivity A permanently drops to* $A' < A$. *On impact, at a given initial capital stock* $K^*(\Gamma, A)$ *investment drops, financial intermediation drops, but financial sector income increases relative to GDP. Over time, the capital stock and intermediation decrease to the new steady state* $K^*(\Gamma, A') < K^*(\Gamma, A)$, *and financial sector income relative to GDP returns to the initial level.*

ii) *Trust* Γ *permanently drops to* $\Gamma' < \Gamma$. *On impact, at a given initial capital stock* $K^*(\Gamma, A)$ *unit investment and financial intermediation drop, and financial sector income decreases relative to GDP. Over time, the capital stock and intermediation gradually fall to the new steady state* $K^*(\Gamma', A) < K^*(\Gamma, A)$, *and financial sector income decreases relative to GDP.*

A drop in either productivity or trust causes financial intermediation to shrink, both in the short and in the long run (at least weakly). In the short run, the two types of shocks entail different responses in the relative size of the financial sector. While a drop in productivity causes the relative size of the financial sector to increase, a drop in trust causes the relative size of the financial sector to decline. This is because the drop in productivity reduces GDP and growth opportunities a lot but

leaves the wealth preservation service of the financial sector relatively unaffected. As a consequence, the financial sector shrinks less than GDP, increasing the share of national income going to finance. In contrast, a drop in trust reduces the remuneration of both the wealth preservation and growth services of the financial sector. Although such a drop also reduces investment and income, on impact it exerts a much more drastic effect on the financial sector income, causing the relative size of finance to drop.

In our model, permanents shocks to productivity or trust can generate long lasting boom and bust cycles to the size of the financial sector. As trust suddenly dissipates (owing, for instance, to a financial crisis), individual take money out of the financial sector and put it into mattresses (self-storage). This reduces financial intermediation and the financing of profitable investment opportunities. Income reductions reduce the stock of wealth, further undermining the ability to finance investment in the future. This process generates a persistent contraction in financial intermediation and income until the new, lower, equilibrium is attained.

Corollary 2 may help make sense of the one dramatic fluctuation in the size of the financial sector in the United States, namely the collapse of its income from 6 to 2 percent of GDP in the great depression, which took 40 years to fully reverse (Figure 1). The Great Depression in all likelihood combined a decline in productivity with a sharp decline in trust in the financial system. Corollary 2 suggests that both of these factors should have led to a progressive decline in the total amount of intermediated wealth. On the other hand, the fact that the income share going to the financial sector immediately shrunk underscore the role of the decline in trust. This is consistent also with the fact that the financial sector started to grow again only after World War II, and reached its prewar size only in the 1980s, decades after the productivity and the wealth of the US economy have substantially surpassed their pre-Depression levels. Only the slow decades-long return of trust enabled the financial sector to reach new heights as the wealth of the US economy expanded.

5. Entry into the Financial Sector

Our analysis has so far focused on the dynamics of fees and of financial intermediaries' income as shaped by the progressive exhaustion of investment opportunities (the diminishing returns assumption). We have so far neglected another important dimension of financial sector evolution, namely entry of new financial intermediaries, which was precluded by the assumption that the number of money managers is fixed at $m = 1/\Delta$.

We now allow for endogenous entry of financial intermediaries. Formally, we allow the distance between two adjacent money managers to change over time. Denote the distance at time t by Δ_t and the corresponding number of financial intermediaries by $m_t = 1/\Delta_t$. For notational simplicity we treat this variable as continuous, even though the number of active intermediaries is equal to the largest integer below m_t. We assume that creating a new money management firm at time t costs $\eta \cdot AK_t^{\alpha}$ units of consumption, where $\eta > 0$. This cost should be viewed as the value of labor that the founder must expend in order to setup the new financial intermediary and to earn the trust of investors (indeed, the opportunity cost of time at t is equal to the wage rate, which scales with value added).[8] Money managers can enter/exit at any time, so current profits are the only determinants of entry decisions. Finally, money managers appear in discrete and thus negligible numbers, so entry of additional managers leaves the labor supply of productive firms unchanged.

Generalizing our previous analysis, equilibrium fees at time t are now given by:

$$f_t^* = \varphi(\Delta_t) \cdot \mathbb{E}(R_t - \gamma), \text{ where } \varphi(\Delta_t) \equiv \left[\frac{\Delta_t}{\Gamma} - \left(\frac{\Delta_t}{2\Gamma}\right)^2\right]. \quad (15)$$

[8] In this formalization, entry entails a redistribution of managers along the unit circle. This redistribution allows all money managers, including the ones entering at t, to be located at the same distance Δ_t. We assume that it is costless for existing managers to relocate along the circle, while it is costly to acquire the necessary generalized trust to operate a firm.

The fee $\varphi(\Delta_t)$ per unit of excess return increases in Δ_t because competition among money managers is less intense when there are fewer managers (Δ_t is higher).

5.1 Entry and Equilibrium Dynamics

If at time t a number $1/\Delta_t$ of money managers is active, the total profits of the financial sector are equal to $f_t^* K_t = \varphi(\Delta_t) \cdot [(1-\gamma)K_t + \alpha K_t^\alpha]$. At time t, money managers enter until the profit earned by each of them is equal to the setup cost. This condition is given by:

$$\frac{f_t^* K_t}{m_t} \equiv \varphi(\Delta_t) \cdot \Delta_t \cdot [(1-\gamma)K_t + \alpha A K_t^\alpha] = \eta \cdot A K_t^\alpha. \tag{16}$$

By dividing both sides by AK_t^α, we can rewrite the equilibrium entry condition as:

$$\varphi(\Delta_t) \cdot \Delta_t \cdot [(1-\gamma)K_t^{1-\alpha} + \alpha] = \eta. \tag{17}$$

Here $\varphi(\Delta_t)$ captures the fee charged by each money manager per unit of service provided (be it wealth preservation or growth). This component increases with Δ_t because a drop in the number of managers raises fees and the aggregate income of each manager.

The second term $\Delta_t \cdot [(1-\gamma)K_t^{1-\alpha} + \alpha]$ on the left hand side captures the share of the aggregate value of money managers' services to aggregate income provided by each individual manager at time t. As shown in the previous section, this ratio increases with the capital stock K_t because financial intermediaries' wealth preservation service becomes relatively more important as the country becomes richer. This feature drives one key property of the entry model, which we summarize in the result below.

Lemma 2 *Consider a path along which the capital stock K_t increases over time. Equation (17) implies that along this path:*

i) *The number of active money managers increases (i.e., Δ_t drops) over time.*

ii) *The management fees charged per unit of capital fall over time, owing both to the drop in $\varphi(\Delta_t)$ as new money managers enter, and to the fall in the marginal return to capital as K_t increases.*

iii) *The aggregate income of the financial sector increases over time, both in absolute terms and relative to the country's aggregate income.*

As the capital stock expands, there are more resources available for intermediation. For given fees, money management becomes more profitable, so incurring the setup cost $\eta \cdot AK_t^{\alpha}$ becomes worthwhile. This stimulates entry of new money managers, leading to a drop in Δ_t until the profits available to an entering money manager drop back to the setup cost. In this process, management fees drop. Part of this effect is due, as in the previous section, to the fact that capital deepening reduces the marginal return on capital and thus the surplus each money manager can extract. Entry, however, introduces another mechanism whereby fees drop: the drop in Δ_t intensifies competition among money managers, reducing equilibrium fees $\varphi(\Delta_t)$.

Despite the drop in unit fees, the aggregate profits of the financial sector increase over time. As before, the expansion in the capital stock increases the demand for financial services. This force, which increases profits, is so strong that it more than offsets the drop in fees. Financial sector income increases not only in absolute terms but also relative to GDP. In equation (17), the left hand side must stay constant, which implies that the total income share absorbed by finance $\varphi(\Delta_t) \cdot [(1-\gamma)K_t^{1-\alpha} + \alpha]$ increases as higher capital stock K_t causes Δ_t to drop.

Lemma 2 considers what happens to entry and to the size of the financial sector as the capital stock K_t grows over time. We still need to verify, however, that with endogenous entry our model delivers an increasing path for the capital stock. In this case, the law of motion of the

economy is still captured by Equations (11) and (12) with the only difference that now also $\varphi(\Delta_t)$ and Δ_t evolve according to Equation (17). In the appendix we then prove the following result.

Proposition 2 *If the parametric conditions of Proposition 1 hold, and in addition productivity A is sufficiently high, the entry model admits a unique and locally stable nonzero steady state* K^*.

Starting from initial levels of capital K_0 below the steady state, the model is indeed capable of generating a transitional growth path characterized by capital deepening, increasing financial intermediation, rising wealth, entry of money managers, decline in fees, but also increasing financial sector income both in absolute terms and relative to GDP. A high level of *A* ensures equilibrium uniqueness by bounding the role of the wealth preservation service provided by the financial sector. If *A* and thus the return from growth services is low, a high capital stock may alone create a strong demand for financial services, generating massive entry of intermediaries in the economy, in turn sustaining massive investment. A large *A* creates a sizeable demand for financial intermediation regardless of the wealth preservation component, precluding the possibility of multiple equilibria.

5.2 Empirical Predictions of The Model with Entry

The possibility of entry of new intermediaries does not affect the ability of our model to account for the main patterns of evolution of the financial sector. As illustrated by Lemma 2, the share of GDP coming from finance increases over time (again due to wealth preservation services), and adverse shocks to investor trust reduce the income going to money managers. The new twist here is that some of these effects now occur through entry of new intermediaries. Most notably, part of the reason why the income of the financial sector grows over time is that the entry of new

intermediaries increases the proximity of money managers to investors, increasing risk taking and the size of the financial sector.

This process of increasing participation into risk taking implies that despite the reduction in the equilibrium unit fee f_t^*, the unit cost of financial intermediation may actually increase as the financial sector expands. To see this, note that the total amount of financial assets in the economy at time t, which includes "safe storage securities" and risky assets is equal to w_{t-1}, the total wealth of the elderly. At the same time, the total income absorbed by the financial sector is equal to the fee times risky investment $f_t^* K_t = f_t^* \theta_t \cdot w_{t-1}$, where θ_t is the wealth share that the elderly allocate to risk taking. The unit cost of financial intermediation is then given by:

$$\frac{f_t^* \theta_t \cdot w_{t-1}}{w_{t-1}} = f_t^* \theta_t.$$

As the financial sector grows, unit fees f_t^* fall but the composition of investment shifts toward riskier assets: θ_t rises. As we show in Appendix B, the latter effect may actually dominate, causing unit costs of intermediation to rise over time.

This idea can help reconcile the Greenwood and Scharfstein (2013) finding that in the last 30 years unit fees have gone down within asset classes with the evidence presented in Figure 5 that the average unit cost of finance – measured as finance income over financial assets – has increased significantly after the Great Depression and then remained roughly constant after 1980 (using a different measure, Philippon 2012 also finds that the average unit cost of finance has slightly increased). The big increase in unit cost after the Depression probably has something to do with recovery of trust and increased risk-taking as the memory of the Depression receded. In addition, the relative stability of unit cost in recent years might be related to the entry of new intermediaries in response to increased wealth, shifting the composition of investments to more intermediated and thus more expensive forms of finance.

These compositional effects may be stronger if money managers can introduce new, riskier, assets. Because these assets rely more on the lubricating role of trust, they may be purchased by investors only if the latter are sufficiently close to their preferred manager. A highly trusted manager would introduce new investment options to leverage his trust capital and extract higher fees.[9] The compositional shift toward higher-risk, nonstandard investments may also constitute a force preventing unit costs from falling over time.

6. Conclusion.

We have presented a Solow-style growth model in which the financial claims on the capital stock are managed by professionals. In that model, the size of the financial sector depends both on the economy's GDP and on its stock of capital or wealth. The model accounts for some key facts about the development of the financial sector in the last century.

At the most basic level, the model explains why financial sector has grown relative to GDP over time (Figure 1 from Philippon 2012). The reason is that one of the functions of finance is to preserve the existing stock of wealth, and wealth has grown over time relative to income, as one would expect along the adjustment path to the steady state. The model thus also predicts the growth of the wealth to GDP ratio over time, shown in Figure 4 and more broadly by Piketty and Zucman (2012).

Equally important, our model's emphasis on trust helps explain aspects of the volatility of the financial sector. Our approach links the sharp decline of finance in the Great Depression, and

[9] To see the logic behind this intuition, note that in our model unit fees increase in the return paid by the financial asset (see Equation (7)). Intuitively, the higher the return earned by investors, the higher the rent that the trusted money manager can extract. As a consequence, the introduction of a higher return-higher risk financial asset by a trusted money manager allows the manager to charge higher fees to investors, increasing the average cost of intermediation. To simplify the analysis, we only consider the choice between a riskless (and thus non-intermediated) asset and a risky asset. We leave the formal analysis of the case of multiple risky assets to future research.

its slow recovery over the following 50 years, to the rapid decline and subsequent slow recovery of trust. Part of that recovery is exogenous, as the memory of the Great Depression recedes, but part of it is also endogenous in our model, since increases in wealth encourage entry by financial intermediaries, which creates high trust relationships.

Our model also seeks to reconcile the somewhat conflicting evidence on the fees and unit costs of the financial sector. On the one hand, there is clear evidence that fees on given financial products have declined over time (e.g., Greenwood and Scharfstein 2013). On the other hand, Philippon (2012) and Figure 5 show no evidence of declining "unit cost" of finance. According to our analysis, an important byproduct of economic growth, entry by financial intermediaries, and reduction in fees is that investors allocate increasing shares of their wealth to intermediated financial products, rather than to self-storage. This implies that the composition of investor portfolios shifts over time to riskier, and hence more expensive, financial products. This can lead to increases in unit costs, even as fees for given products decline.

Previous scholars took the evidence on the growth of finance as an indication of problems with the market economy and the financial system. Without denying the importance of rent-seeking, agency, and other problems, our paper presents a more benign view. Finance *should* grow as an economy matures, because the preservation of wealth is an increasingly important function of the financial system.

References

Baumol, William. 1967. "Macroeconomics of Unbalanced Growth: The Anatomy of the Urban Crisis," *American Economic Review* 57(3): 415–26.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2012. "Money Doctors," Working paper.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2004. "The Role of Social Capital in Financial Development," *American Economic Review* 94(3): 526–556.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2008. "Trusting the stock market," *Journal of Finance* 63, 2557–2600.

Greenwood, Robin and David Scharfstein. 2013. "The Growth of Finance," *Journal of Economic Perspectives* 27(2): 3–28.

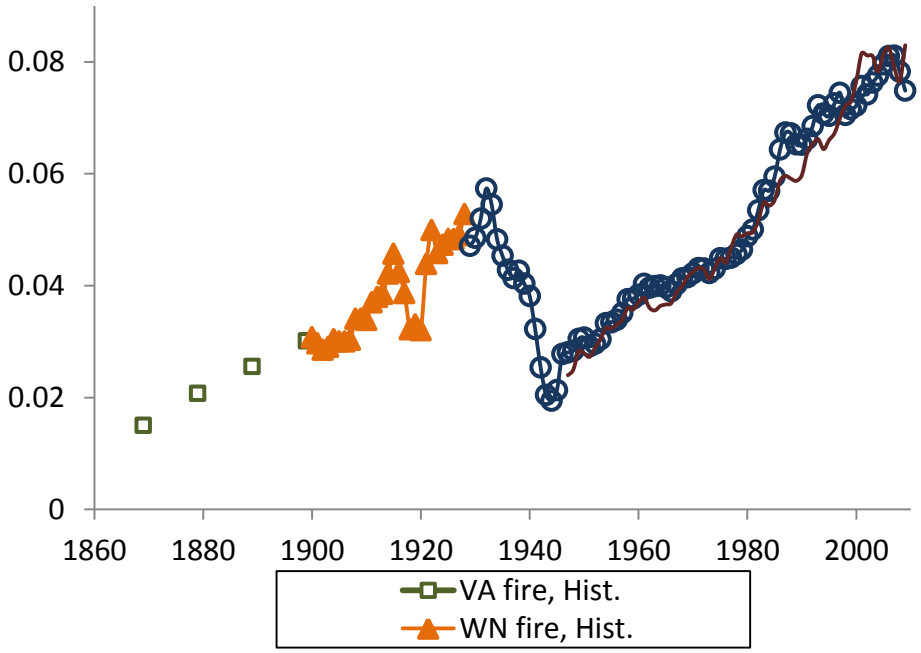
Philippon, Thomas. 2012. "Has the U.S. Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation," Working paper.

Philippon, Thomas and Ariell Reshef. 2012. "Wages and Human Capital in the U.S. Financial Industry: 1909–2006," *Quarterly Journal of Economics* 127(4): 1551–1609.

Philippon, Thomas and Ariell Reshef. 2013. "An International Look at the Growth of Modern Finance," *Journal of Economic Perspectives* 27(2): 73–96.

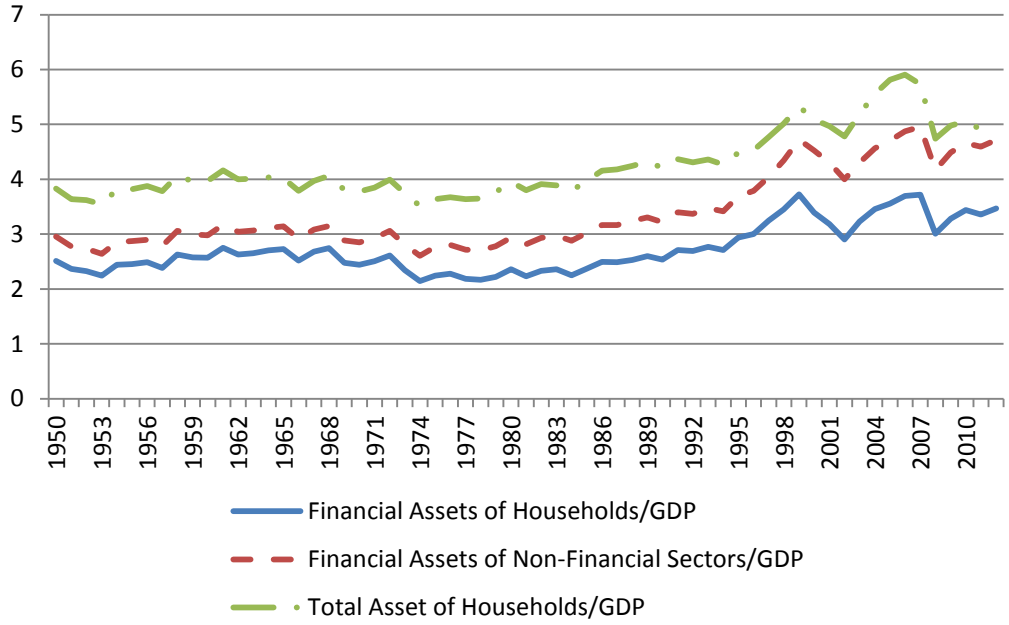
Piketty, Thomas and Gabriel Zucman. 2012. "Capital is Back: Wealth-Income Ratios in Rich Countries 1870-2010," Presentation.

Figure 1. Financial Sector Income/GDP



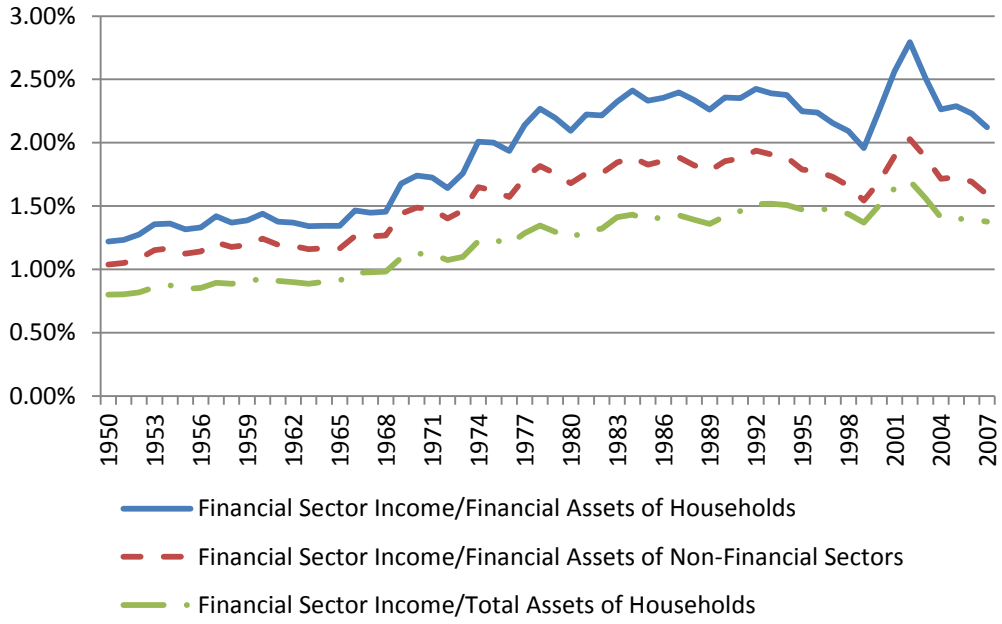
Notes: VA is value added, WN is compensation of employees, "fin" means finance and insurance, "fire" means finance, insurance, and real estate. For "NIPA", the data source is the BEA, and for "Hist" the source is the Historical Statistics of the United States. Directly from Philippon (2012).

Figure 4. Financial Assets/GDP



Source: Philippon (2012) and Flow of Funds. Non-financial sectors include households, nonfarm businesses.

Figure 5. Financial Sector Income/Financial Assets



Source: Philippon (2012) and Flow of Funds. Non-financial sectors include households, nonfarm businesses.

Appendix A: Proofs

Proof of Proposition 1. By plugging Equation (12) into (11), it is easy to see that in any steady state with positive capital stock $K^* > 0$ and such that $\theta^* < 1$, is identified by the equation:

$$K^* = \frac{(1-\varphi)(1+\alpha A(K^*)^{\alpha-1}-\gamma)}{\sigma[1+A(K^*)^{\alpha-1}]^2} \cdot \left(\Gamma - \frac{\Delta}{4}\right) \cdot (1-\alpha)A(K^*)^{\alpha},$$

which can be rewritten as:

$$c \cdot [(K^*)^{1-\alpha} + A]^2 = \left[(1-\gamma)(K^*)^{(1-\alpha)} + \alpha A\right] \cdot A, \qquad (P1)$$

where $c \equiv \frac{\sigma}{(1-\varphi)\cdot\left(\Gamma-\frac{\Delta}{4}\right)\cdot(1-\alpha)}$. It is easy to verify the above equation admits a unique solution $K^* > 0$ provided $c < \alpha$, which imposes an upper bound on σ.

Before studying the steady state, let us verify that $\theta^* < 1$ (and in particular that this is so for all investors). From Equation (12) it is easy to find that the household who is closest to a manager invests a share of wealth:

$$\theta(K_t^{1-\alpha}) = \frac{K_t^{1-\alpha}[(1-\gamma)K_t^{1-\alpha} + \alpha A]}{c \cdot [K_t^{1-\alpha} + A]^2} \cdot z,$$

where $z = \frac{\Gamma}{\left(\Gamma-\frac{\Delta}{4}\right)\cdot(1-\alpha)}$. The function $\theta(\cdot)$ is increasing in $K_t^{1-\alpha}$ provided $K_t^{1-\alpha} < A$, which – as we will soon see, it is strictly satisfied at the steady state capital level, and thus along transitional dynamics occurring around the steady state. which implies that starting from a below steady state level of capital stock, risk taking increases over time until the steady state is reached. As a result, by exploiting Equation $(P1)$, all investors set an interior level of risk taking in the steady state provided $(K^*)^{1-\alpha} < A/z$, where $z > 1$. By replacing this condition into $(P1)$ we find that this is equivalent to:

$$c > \frac{z \cdot [(1-\gamma) + \alpha z]}{(1+z)^2},$$

which imposes a lower bound on σ. The upper and lower bounds are mutually compatible, namely $\frac{z\cdot[(1-\gamma)+\alpha z]}{(1+z)^2} < \alpha$, provided $2\alpha > (1-\gamma)$, which we assume to hold. This analysis thus identifies variance bounds $\bar{\sigma}$ and $\underline{\sigma}$, with $\bar{\sigma} > \underline{\sigma}$, to which we restrict the analysis of our model.

Consider the steady state prevailing for $\sigma \in (\underline{\sigma}, \bar{\sigma})$. This is identified by Equation $(P1)$. By applying the implicit function theorem, and after some algebra, one can find that:

$$\frac{d(K^*)^{1-\alpha}}{dA} \propto -\frac{-c(K^*)^{2(1-\alpha)} + cA^2 - \alpha A^2}{2c[(K^*)^{1-\alpha} + A] - (1-\gamma)A} > 0, \qquad (P2)$$

$$\frac{d(K^*)^{1-\alpha}}{dc} \propto -\frac{[(K^*)^{1-\alpha} + A]^2}{2c[(K^*)^{1-\alpha} + A] - (1-\gamma)A} < 0, \qquad (P3)$$

where both inequalities rely on the restriction $(K^*)^{1-\alpha} < A/z$ and $c < \alpha$. Condition (P2) intuitively says that the steady state capital stock increases in productivity A. Condition (P3) says that the steady state capital stock increases in the number of managers (because lower Δ reduces c).

Consider now the dynamics of the model. By exploiting Equations (11) and (12), one can write the law of motion for our model economy as:

$$K_t^{\alpha} \frac{(K_t^{1-\alpha} + A)^2}{[(1-\gamma)K_t^{1-\alpha} + \alpha A]} - \frac{1}{c} A K_{t-1}^{\alpha} = 0. \qquad (P4)$$

The above difference equation implicitly defines a function $K_t(K_{t-1})$ whose slope is equal to:

$$\frac{dK_t}{dK_{t-1}} = \frac{\frac{1}{c} \cdot \alpha A}{K_{t-1}^{1-\alpha} \cdot \frac{(K_t^{1-\alpha} + A)^2}{[(1-\gamma)K_t^{1-\alpha} + \alpha A]} \left\{ \frac{\alpha}{K_t^{1-\alpha}} + \left(\frac{1-\alpha}{K_t^{1-\alpha} + A} \right) \frac{[(1-\gamma)K_t^{1-\alpha} + (\alpha - 1 + \gamma)A]}{[(1-\gamma)K_t^{1-\alpha} + \alpha A]} \right\}}.$$

At the $K_t = K_{t-1} = 0$ steady state, the above slope becomes equal to:

$$\frac{dK_t}{dK_{t-1}} = \frac{\alpha}{c} > 1,$$

Where the inequality is due to the assumption $c < \alpha$. As a result, the zero capital steady state is unstable, and the mapping $K_t(K_{t-1})$ must cut the 45 degrees line at the interior steady state K^* with a slope less than one, implying that K^* is locally stable.

Proof of Corollary 2 At the steady state capital sock $K^*(\Gamma, A)$, the new productivity level A' sets the wage rate, fees and intermediation at time t. In particular, investment and intermediation are pinned down by the equations:

$$K_{t+1} = \theta_{t+1} A \left(\frac{w_t}{A}\right),$$

$$\theta_{t+1}(K_{t+1}, A) = \frac{(1-\varphi)(1 + \alpha A K_{t+1}^{\alpha-1} - \gamma)}{\sigma[1 + A K_{t+1}^{\alpha-1}]^2} \cdot \left(\Gamma - \frac{\Delta}{4}\right).$$

where $\left(\frac{w_t}{A}\right)$ is by definition invariant to changes in A, for the initial capital stock is predetermined.

Consider the effects of a change in A. The impact of such change on investment and intermediation is determined by the behavior of the ratio $K_{t+1}/\theta_{t+1}(K_{t+1}, A)A$. By the proof of proposition 1 we know that such ratio is an increasing function of K_{t+1} and a decreasing function of A at the steady state capital level. As a result, by the implicit function theorem, a drop in productivity reduces financial intermediation and the capital stock K_{t+1}. The relative size of the financial sector depends on the effect of the productivity change on the product $AK_{t+1}^{\alpha-1}$. Denote $x \equiv K_{t+1}^{1-\alpha}/A$. The relatie size of finance increases with x. In this regard, note that the equilibrium condition $\frac{K_{t+1}}{\theta_{t+1}(K_{t+1}, A)A} = M$, where M is a constant, can be rewritten as:

$$A^{\alpha} \frac{x}{[\theta_{t+1}(1/x)]^{1-\alpha}} = M.$$

After some algebra, one can check that the left hand side of the above equation increases in x. As a result, an increase in A reduces x and thus the relative size of the financial sector, while a drop in A does the reverse. Finally, consider the long run response. It is easy to see from the Proof of Proposition 1 and from Equation (P1), financial intermediation drops in the long run and the relative size of the financial sector remains constant.

Consider now the effect of a change in trust Γ. The equilibrium condition is the same as the one represented above. Because the function $\theta_{t+1}(K_{t+1}, \Gamma)$ increases in Γ, higher trust increases investment and intermediation, while a drop in trust does the reverse. Accordingly, because also the function $\theta_{t+1}(1/x, \Gamma)$ increases in Γ, an increase in trust on impact increases the relative size of the financial sector while a reduction in trust does the reverse. Finally, in the Proof of Proposition 1 we also establish that long run intermediation and the long run relative size of finance increase in trust.

Proof of Lemma 2. By inspection of Equations (16) and (17).

Proof of Proposition 2 With endogenous entry, the evolution of the economy is described by the following equations:

$$K_t = \frac{(1 + \alpha A K_t^{\alpha-1} - \gamma)}{\sigma[1 + A K_t^{\alpha-1}]^2} \cdot (1 - \varphi_t) \cdot \left(\Gamma - \frac{\Delta_t}{4}\right) \cdot (1 - \alpha) A K_{t-1}^{\alpha}, \qquad (P5)$$

$$\Delta_t \cdot \varphi(\Delta_t) \cdot \left[\frac{(1-\gamma)}{A} K_t^{1-\alpha} + \alpha\right] = \eta. \qquad (P6)$$

Equation $(P5)$ is essentially the same law of motion of the Proof of Proposition 1, with the only difference that now Δ_t (and thus φ_t) are endogenously determined in Equation $(P6)$.

In the spirit of the Proof of Proposition 1, we can rewrite (P5) as:

$$K_t^{\alpha} \frac{\sigma[K_t^{1-\alpha} + A]^2}{\left[\frac{(1-\gamma)}{A} K_t^{1-\alpha} + \alpha\right]} \cdot \frac{1}{(1 - \varphi_t) \cdot \left(\Gamma - \frac{\Delta_t}{4}\right)} = (1 - \alpha) A^2 K_{t-1}^{\alpha}. \qquad (P7)$$

By replacing in Equation $(P6)$ the expression for $\varphi(\Delta_t)$ and by denoting $s(x) \equiv \left[\frac{(1-\gamma)}{A} x + \alpha\right]$, we can find after some algebra that:

$$\left(\frac{\Delta_t}{\Gamma}\right)^2 - \frac{1}{4}\left(\frac{\Delta_t}{\Gamma}\right)^3 = \left[\frac{\eta}{\Gamma s(x)}\right],$$

Where $x \equiv K_t^{1-\alpha}$. This equation has a unique solution for Δ_t/Γ in (0,1) which we denote by $\psi(x)$.

By replacing the expression for $\psi(x)$ in the expressions for φ_t and Δ_t in Equation $(P7)$, we find after some algebra that the law of motion of the economy is given by:

$$K_t^{\alpha} \cdot \frac{\sigma[x + A]^2}{\Gamma s(x) \left[1 - \psi(x) + \frac{\psi(x)^2}{4}\right] \cdot \left[1 - \frac{\psi(x)}{4}\right]} = (1 - \alpha) A^2 K_{t-1}^{\alpha},$$

Where again we have that $x \equiv K_t^{1-\alpha}$. The above difference equation has one trivial steady state at $K_t = x = 0$. A positive and unique steady state exists provided: i) the root multiplying K_t^{α} on the left hand side above is monotonically increasing in x, ii) the value of the root at $x = 0$ is below $(1-\alpha)A^2$. The latter condition is met when the variance σ is sufficiently low. On the other hand, a sufficient condition for i) is that:

$$s'(x) = \frac{(1-\gamma)}{A} \quad \text{is sufficiently small.}$$

Intuitively, in this case the main effect of higher x is to increase the numerator, leaving the denominator almost unaffected (also because in this case $\psi'(x)$ stays small). When this is the case, then, there is a unique interior equilibrium $K^* > 0$. This equilibrium is locally stable (so that the capital stock monotonically converges to it) provided the slope of the implicit mapping $K_t(K_{t-1})$ is above one at the $K^* = 0$ steady state. One can check that this is the case provided A is sufficiently high and σ is above a threshold (consistent with the previous upper bound). The condition that σ be bounded is the same as the one required in Proposition 1, except that now the bounds are evaluated at the equilibrium number of managers prevailing when $x = 0$ as entailed by $\psi(0)$. Since $\psi(0)$ does not depend on productivity A, the assumption that A be sufficiently large can be added to ensure stability of the system. Note that when $\psi'(0)$ is made small, the upper and lower bound will be consistent because locally entry responds slowly to changes in the capital stock, so that around $x = 0$ the analysis does not virtually change from that with a fixed number of money managers.

Appendix B: Extensions and Additional Proofs.

B.1 Population Growth and Technological Progress

In the main model, we have assumed that population – and thus labor supply – is constant at $L_t = 1$ and that total factor productivity takes a constant value of A. We now allow for population growth and for productivity augmenting technological progress and investigate how these features affect the evolution of the financial sector. We do so by assuming that the effective labor supply available at time t satifies the law of motion:

$$L_t = (1+n)(1+x)L_{t-1},$$

where n is the rate of population growth and x is the rate of technical progress. Because the production function is Cobb-Douglas, this formulation of labor augmenting technical progress is equivalent to one in which productivity growth is factor-neutral and increases the value of A.

Denoting by $\widehat{K}_t \equiv K_t/L_t$ the capital stock per unit of effective labor, the competitive remunerations of a unit of effective labor and of a unit of capital are respectively given by:

$$(1-\alpha)A\widehat{K}_t^{\alpha} = w_t,$$

$$R_t = \varepsilon_t + \varepsilon_t \cdot A \cdot \widehat{K}_t^{\alpha-1} - (1-\alpha)A\widehat{K}_t^{\alpha-1},$$

where the second expression implies, consistent with our previous analysis, that the average return on a unit of capital is equal to $\mathbb{E}\{R_t\} = 1 + \alpha A\widehat{K}_t^{\alpha-1}$, while the variance of the return to capital is equal to $\sigma_t = var(R_t) = \sigma\left[1 + A\widehat{K}_t^{\alpha-1}\right]^2$.

Thus far, the only implication of introducing population growth and technical progress is that factor remunerations are functions of the capital per effective unit of labor $\widehat{K}_t$. This implies that the share of wage income invested into risky asset also depends in $\widehat{K}_t$, namely:

$$\theta_t = \frac{(1-\varphi)\left(1+\alpha A\widehat{K}_t^{\alpha-1}-\gamma\right)}{\sigma\left[1+A\widehat{K}_t^{\alpha-1}\right]^2}\cdot\left(\Gamma-\frac{\Delta}{4}\right).$$

The absolute value K_t of the capital stock created at time t is then equal to a fraction θ_t of the total wage bill paid to workers at $t-1$, which is equal to $w_{t-1}\cdot L_{t-1}$, namely $K_t = \theta_t \cdot w_{t-1} \cdot L_{t-1}$. By dividing both sides by L_t, and by replacing in the equation the expression for w_{t-1}, we find that the law of motion of the capital stock per unit of effective labor is given by:

$$\widehat{K}_t = \frac{\theta_t}{(1+n)(1+x)}\cdot(1-\alpha)A\widehat{K}_{t-1}^{\alpha}.$$

The law of motion characterizing the capital stock per unit of effective labor is very similar to that described in Equation (11) of our basic model, except that the fraction of wealth invested in the risky asset is scaled down by the constant population and productivity growth rates.

In light of the previous analysis, several immediate consequences follow. First, the capital stock per unit of effective labor converges to a nonzero steady state value $\widehat{K}^*$ that is a decreasing function of n and x. In this steady state, the per-capita capital stock and per capita output grow at a constant rate x, while the extent of risk taking θ_t converges to a constant. The comparative statics properties described by Proposition 1 continue to hold with respect to the steady state levels of the per capita capital stock and of the extent of risk taking.

Second, the properties of evolution of the financial sector also do not change from Corollary 1. The management fee per unit of capital declines over time as $\widehat{K}_t$ increases toward its steady state level. As a consequence, financial sector income rises faster than value added if we express both the numerator and the denominator in per effective units of labor.

Finally, the qualitative properties of Corollary 2 also hold in this modified model. In sum, population and productivity growth introduce additional reasons for the growth of the absolute size and profits of the financial sector, but do not affect the qualitative behavior of scaled variables such as unit fees and the income share going to finance.

B.2 Trading and Valuation of the Capital Stock

In our baseline model consumption and capital are the same good, so that the elderly consume the capital stock they own at the end of their lives. This assumption simplifies the analysis, but it raises the issue of whether our result are robust to the more realistic setting in which capital cannot be converted back into consumption and so the elderly must sell their capital stock to the young. To shed light on this issue, suppose now that the consumption can be transformed into capital but capital cannot be converted back into consumption. This implies that at time t the elderly of the generation born at time $t-1$ must sell the economy's capital stock to the current young generation. The amount of capital held by the elderly at the end of time t is equal to $\varepsilon_t \cdot K_t$. If the price of capital in terms of consumption is p_t, the value at time t of the supply of capital in terms of consumption goods is equal to $p_t \cdot \varepsilon_t \cdot K_t$. On the demand side, the consumption income available to the young born at time t to buy – through money managers – the entire capital stock from the elderly is equal to $\theta_{t+1}\cdot w_t$. Of course, the young only demand capital from the elderly if

the price of existing capital is not higher than the resource cost of creating new capital, i.e. provided $p_t \leq 1$, which importantly affects equilibrium prices.

To find the equilibrium price p_t, we must determine whether the capital stock $\varepsilon_t \cdot K_t$ available at time t is below or above the desired investment $\theta_{t+1} \cdot w_t$ by the young born at t. If the young wish to increase the stock of capital, namely $\varepsilon_t \cdot K_t < \theta_{t+1} \cdot w_t$, the equilibrium price of capital settles at $p_t = 1$ so as to make savers indifferent between buying existing capital goods and creating new ones. If instead the young wish to reduce the stock of capital, namely $\varepsilon_t \cdot K_t > \theta_{t+1} \cdot w_t$, then the new capital goods will not be produced and the price drops to $p_t = \frac{\theta_{t+1} \cdot w_t}{\varepsilon_t \cdot K_t} < 1$ so as to equate the values of the demand and the supply of capital goods.

Because our main results focus on transitions occurring below the steady state, let us consider the implications of this analysis for changes in the valuation of capital markets during these transitions. Recall that in these transitions, the desired capital stock increases over time, namely $K_{t+1} = \theta_{t+1} \cdot w_t > K_t$. As a consequence, if the potential shocks ε_t are sufficiently small that below the steady state capital the condition $\varepsilon_t \cdot K_t < \theta_{t+1} \cdot w_t$ holds (at least when K_t is far enough from the steady state), then during the transitional growth phase the unit price of capital stays constant at $p_t = 1$. In each period, the elderly sell their capital $\varepsilon_t \cdot K_t$ to the young, who add extra investment to implement their desired capital stock $\theta_{t+1} \cdot w_t$. The ex-post shock ε_t affects consumption by the elderly and new investment by the young, but leaves the aggregate capital stock next period unaffected. The law of motion of the economy is then identical to Equation (11): the possibility to trade capital goods does not affect how the economy converges to the steady state.

The possibility of trading in capital goods, however, affects the interpretation of our results. In particular, the capital stock K_t can now be interpreted as the market valuation of the aggregate wealth of the economy. The fact that the income share of the financial sector raises with K_t can then be viewed as the product of increasing capital market valuations. It should be noted, however, that in our model these valuations rise through the extensive margin – as new investment takes place – and not through increases in their unitary valuation p_t, which remains constant at 1. Allowing for changes in p_t, potentially through asset price bubbles, is an interesting avenue for future research.

B.3: Competitive Entry of Intermediaries and the Growth of Financial Sector Income

We now show that it is possible that the unit cost of finance (the ratio of financial sector income over financial assets):

$$f_t^* \theta_t = \varphi_t(\Delta_t) \cdot (1 - \varphi_t(\Delta_t)) \cdot \left(\Gamma - \frac{\Delta_t}{4}\right) \cdot \frac{(1 + \alpha A K_t^{\alpha-1} - \gamma)^2}{\sigma[1 + A K_t^{\alpha-1}]^2},$$

may increase over time, as new intermediaries enter the market. To see why this may be the case, note that during transitional growth, the capital stock K_t increases while the distance between managers Δ_t decreases. As a result, a sufficient condition for the product $f_t^* \theta_t$ to increase over time is that the terms that are functions of Δ_t decrease in Δ_t while ratio which is a function of K_t increases in K_t. It is immediate to see that the ratio on the right increases in K_t provided $\alpha < 1 - \gamma$. On the other hand, one can find values such that the first term (which is a polynomial of degree 5) decreases in Δ_t (e.g. Δ_t close to Γ). It is beyond the scope of this analysis to evaluate under what exact conditions unit costs may be increasing, but it seems that – given that Δ_t is pinned down by η – one may be able to find economies (values of η and of the initial capital stock) for which the equilibrium Δ_t is indeed close to Γ and unit costs increase over time until the steady state is reached.