Competition, Common Agency, and the Need for Financial Intermediation

Anton van Boxtel[*]

University of Vienna, Vienna Graduate School of Finance

August 20, 2021

Abstract

This paper argues that financial intermediaries serve to overcome competitive externalities between investors. A borrower has access to a long-term project that is subject to an uncertain intermediate date refinancing need. Investors compete to offer financing contracts but cannot control or verify the firm's contracts with other investors. If the firm deals with each investor bilaterally, this leads to a double common agency problem: on the one hand it becomes difficult to limit the maximum refinancing ex ante. On the other hand, when multiple investors contract with the firm at the same time, they want to make sure the other investors supply the refinancing. An intermediary arises naturally to bundle various investors' resources and unilaterally deal with the borrower. This can explain why certain banking models, such as universal banking, investment banking, or syndicated lending emerged as the way to finance long-term industrial investment with large uncertain refinancing needs.

Keywords: Non-exclusivity, common agency, financial intermediation, credit markets.
JEL Classifications: D21, D82, G21.

1 Introduction

Banks, and other financial intermediaries, play an important role in bringing capital from investors to entrepreneurs and households in any advanced economy around the world. The reason why banks are needed as a middle-man has been debated by economists. This paper aims to provide a novel rationale for the existence of bank-like financial intermediaries:

[*] Email: aavanboxtel@gmail.com; I am very thankful to Andrea Attar, Fabio Braggion, Fabio Castiglionesi, Fabio Feriozzi, Marco Da Rin, and Klaus Ritzberger, as well as seminar participants at the Institute for Advanced Studies in Vienna, and conference participants at the 2019 Finance Forum in Madrid for their comments and support.

banks arise as an institution to coordinate competition between investors for long term investment.

In this paper, intermediaries are necessary as an institution to coordinate the flow of capital from investors competing in a *non-exclusive* and *uncoordinated* fashion to borrowers. Competition is non-exclusive in the sense that investors can neither observe nor control the contracts that borrowers have with investors at later dates. Competition is uncoordinated in the sense that if several investors finance the same borrower together, they cannot collude on the contracts that each one of them trades with that borrower.

This paper features a Holmström and Tirole (1998) type model of investment: borrowers have access to a project that requires an initial fixed investment, pays off at a later date and is subject to a liquidity shock at an intermediate date. Under optimal contracting, investors provide money to the borrower for the initial investment and set a maximum level up to which the liquidity shock can be financed at the intermediate date. In exchange for their investment, they ask for a repayment from the project's proceeds at the final date. As the project is subject to moral hazard, this repayment needs to be limited.

Because of the non-exclusive and uncoordinated nature of the contracting, two distinct, and converse, free-riding problems arise. On the one hand, a similar problem as in Boxtel, Castiglionesi, and Feriozzi (2013) is present: for any finite credit line that incumbent investors provide, outside investors can free-ride upon this provision and offer an additional "emergency" liquidity provision. They can do so by obtaining a repayment from the borrower's incentive share. This paper finds that in many cases, it is optimal to preempt such free-riding behaviour by providing liquidity in all states of the world.

On the other hand, if liquidity support is provided in all states of the world, various investors have to provide liquidity together, as the shock might exceed the limited endowment of each individual investor. The potential expected repayment that the borrower can offer remains limited. This means each investor has an incentive to change the pricing of her liquidity provision in such a way that the other investors are responsible for providing liquidity, thus investing and not insuring, free-riding on the insurance others are providing.

The only way the optimal policy of unlimited liquidity support can be implemented, is by the various investors depositing their endowments with an intermediary, and the intermediary offering a contract with full liquidity insurance. As this maximizes surplus for the borrower, an intermediary will arise endogenously in the context of a competitive market.

One investor becomes the intermediary for others and thus the sole entity directly trading with a company. This could explain various trends in the history of financial interemediation. First, it could explain the importance of universal banks in the latter half of the nineteenth century, especially in Germany[1]. Large (universal) banks have played

[1] In the literature review I will give more references to papers discussing the role of bank financing in

an important role in the financing mining, railroads, utilities, and heavy industries. This paper models production technologies that require some initial investment and face potentially high, yet initially uncertain, costs or reinvestment needs at a later date, precisely the salient features of technologies that played a role in these industries: after investment, development costs are still uncertain and in case of a technical failure or unexpected cost overrun, it is next to impossible to scale down operations.

Second, this paper could shed light on consolidation and concentration in the banking sector. The literature documents merger waves in the 1990s in the US (Calomiris, 1999; Calomiris and Karceski, 2000) and in the 1980s and 1990s in Europe (Karceski, Ongena, and Smith, 2005; Boot, 1999). This paper could shed light on how increased international and domestic competition, new technologies, and deregulation might have precipitated these concentration movements: banks needed to pool their investment together, to circumvent the non-exclusivity and common agency problems arising from many smaller banks competing with one another.

Third, a banking sector with multiple banks competing should develop "intermediaries for intermediaries" for financing large-scale long-term projects. In this way, the model in this paper can be applied to understand the origin of investment banking, and, more recently, the rise of syndication. In most syndicated deals, a group of banks appoint one bank as the *lead arranger*, who is also in charge of credit lines to the firm receiving the loan. This lead bank thus becomes an incidental intermediary for the others in a dynamic not dissimilar to the one described in this paper.

The rest of the paper is set up as follows: section 2 reviews the extant historical and empirical evidence, and discusses related theoretical literature. Section 3 describes the model and the underlying assumptions. Section 4 derives the second best and third best allocations and shows how the latter can feature full insurance. Section 5 contains the main result: under full insurance intermediaries arise as a financing form. Section 6 compares intermediation to other forms of financing in the model. Section 7 studies the robustness of results to the various underlying assumptions of the model.

2 Related Literature

This paper models financial intermediaries as a means to overcome two types of externalities between investors. To make a very blunt classification, these sorts of externalities fall within the type commonly studied in the *common agency* literature on the one hand, and the *non-exclusivity* literature on the other. Even though these two literatures are very similar and closely related at a theoretical level, the common agency literature tends to deal with situations in which an agent finds himself contracting with multiple princi-

industrial development

pals, whereas the non-exclusivity literature deals with situations in which the agent *could* contract with several principals.

In the classic literature on competition with asymmetric information it is often, implicitly or explicitly, assumed that agents only deal with only one of many principals. The competitive mechanism of principals undercutting each others' offers would then lead to a solution that is optimal for the agent, only constrained by the relevant moral hazard or adverse selection problems. However, as Pauly (1974) notes, the possibility of agents privately contracting with several counterparties at the same time leads to equilibria that are inefficient also with respect to the constrained optima, as those constrained optima can leave room for a private trade between the agent and non-incumbent principals. These trades would then impose externalities on incumbent lenders. Following this principle, (Bizer and DeMarzo, 1992) model a consumption economy in which a borrower can sequentially approach multiple banks. Under the constrained optimal contract, the borrower has an incentive to approach other banks.

This paper aims to address the formation of large bank-like intermediaries as the most common instrument in an economy to get funds from investors to firms. The economic history literature has noted some times and places where banks started playing an exceptionally large role. Especially in Germany in the late 19th and early and 20th centuries, large universal banks dominated the financing of German firms. The special relation between German banks and industrial firms has been noted by contemporaries (Jeidels, 1905; Riesser, 1910). In a seminal analysis, Gerschenkron et al. (1962) states that

> ...the German banks, and along with them the Austrian and Italian banks, established the closest possible relations with industrial enterprises. A German bank, as the saying went, accompanied an industrial enterprise (...) throughout all the vicissitudes of its existence.

He argues that it was the presence of large universal banks, or *Großbanken* that allowed the German economy to mobilize enough capital for the second industrial revolution and to "catch up" with the more industrialized economy of the United Kingdom.

Numerous studies have compared the German experience to the experience in other countries. During the American "Gilded Age" a number of large financiers, the best known of which is J.P. Morgan, have played a pivotal role in financing American industrialization. The financiers of the house of Morgan were often active on the boards of directors of the firms they financed. DeLong (1991) finds that firms with J.P. Morgan representatives on their boards were 20% more valuable than those without. Ramirez (1995) finds that this difference is most likely attributable to liquidity issues. Having close ties to a bank makes it easier for firms to raise funds in times of high liquidity needs. An interesting case study is presented in Chandler (1954), focusing on the patterns of railroad financing in the US:

even though railroads were often equity financed, firms relied on intermediaries to raise equity. Often larger equity-financed railroads ran into liquidity problems, as one would expect from the analysis in this paper.

Calomiris (1993, 1995) compares the German system to the American one and argues that the regulatory branching and activity restrictions on banks were such that the American system performed significantly worse in financing industrial development. The other comparison that is often made is between the United States and Germany on the one hand, where bank financing played a relatively important role and Great Britain on the other, where, according to Gerschenkron et al. (1962), banks were "obsessive about liquidity and only lent on a short term, hands-off basis." (cf. Guinnane, 2002). Davis (1963) hypothesizes that industrial development in the U.K. in the late nineteenth century started lagging behind that in the U.S. and Germany, because the U.K. lacked the kind of large financial institutions that the U.S. and Germany had. This difference in growth has been documented extensively in Lewis (1978) and attributed by some scholars (such as DeLong, 1991) to the different financial systems present in the different countries.

A few papers compare financing by universal banks to other forms of financing *within* an economy. The aforementioned papers by DeLong (1991) and Ramirez (1995) do so for the United States around the turn of the twentieth century. Hoshi, Kashyap, and Scharfstein (1991) compare firms in post-WWII Japan that have close ties to so-called *keiretsus*, large financial conglomerates, to firms that lack these close ties. They find that the former group of firms has a smaller sensitivity to liquidity shocks, indicating that they are less liquidity constrained. Becht and Ramírez (2003) perform a similar exercise, looking at bank affiliation in Germany and find that mining and steel companies with ties to universal banks were significantly less liquidity constrained than non-affiliated firms.

This paper is close to Diamond (1984) in the sense that it models the necessity of banks from features of the borrowers' investment technology. However, this paper is substantially different from Diamond (1984). It is the inability of both agents and principals to commit to contracts that causes inefficiencies in this model. The technologies in this paper are exogenous, so a monitoring problem as in Diamond (1984) does not exist. Investors have no special ability to overcome information asymmetries that exist between investors and borrowers, be it through some costly monitoring expenditure or through learning. In Diamond (1984), banks perform the economically productive task of monitoring, and centralizing this task to one party gives economies of scale. In this paper, however, intermediaries perform no productive task and are merely needed to overcome contractual externalities between investors.

Holmström and Tirole (1998) argue that in the absence of cash or another storage technology, intermediaries can restore the productive optimum because they can commit to long-term financing in a way that individual investors cannot. This argument is different from the one presented in this paper: intermediaries arise even if a storage technology is present, and the intermediary has no superior commitment power, just the capacity to bundle funds.

Another paper close to this one is the paper by Dewatripont and Maskin (1995), that studies the trade-off between centralized bank financing and decentralized market financing in a model with refinancing: banks might be inclined to refinance too often and decentralized financiers do not refinance often enough. This takes place in a world where financiers cannot commit ex ante to refinancing or not, and where refinancing costs are public information. In the current paper the reasoning is reversed: both under- and overinsurance can be part of a constrained optimal allocation, and financing by either banks or multiple parties arises endogenously as a means to implement each respective allocation.

3 Model

The model is an adaptation of Holmström and Tirole (1998): there are three dates, $t = 0, 1, 2$, and a single good, called money. There is a single *borrower* and a large number M of *investors*. The investors are indexed by $i = 1, 2, \ldots, M$ and each investor has the same endowment W. The assumption that there are multiple investors, but only one borrower, guarantees that if contracting is exclusive, investors make zero profits and maximize the borrower's surplus.

Borrower and Projects The borrower has access to a project that requires an input of money at two different dates: at $t = 0$ an initial investment is needed and at $t = 1$ there is a *liquidity shock*. The borrower has its own endowment A and at $t = 0$ borrows $I - A$ from investors in order to realize an investment of an endogenously determined *size* I. The borrower is protected by limited liability.

The liquidity shock is an exogenous stochastic cost that realizes at $t = 1$ and that needs to be financed in order for the borrower to be able to continue the project. This shock can be thought of as a repair cost after a technical failure or an investment that is needed and of which the costs weren't certain at the inception of the project. If a firm cannot pay this liquidity shock, it cannot continue the project, i.e. there is no possibility of scaling down if only part of the necessary funds can be raised. This assumption is not unrealistic in heavy industries, where a technical failure of one reactor, smelter, or blast furnace can shut down an entire production process. For a mining project or a railroad, one could think of natural obstacles that have to be cleared in order for the project to begin operation.

This liquidity shock is modelled as a dimensionless random variable ϑ. For simplicity, it is assumed that ϑ can take three values, ϑ_s for $s \in \{l, m, h\}$ with states s occurring with respective probabilities $f(\vartheta_s)$. For an initial investment size I, and a shock ϑ, the borrower needs to raise ϑI. If he can, he will do so in order to finance the project and if he cannot, the project has to be abandoned and will yield zero. Following Holmström and Tirole (1998), I assume that there are no investment or consumption opportunities at $t = 1$, and that the borrower cannot divert cash. This means any money withdrawn in excess of ϑI is wasted. The only possible thing the borrower can do with excess liquidity is "burn" it, so that it only matters for the borrower whether the total amount of cash he can raise is larger than ϑI or not.

Moral hazard If the project is continued, the firm decides on an effort level, which can be either low ($e = L$) or high ($e = H$). If the borrower chooses low effort (or *shirks*), he will obtain a non-transferrable private benefit $B > 0$. If he chooses high effort, there is no private benefit. The effort level also determines the success probability p_e of the project, with $p_H > p_L > 0$. The difference between the two probabilities is called $\Delta p := p_H - p_L$. If the project is successful, it yields a pecuniary return R. The expected return given continuation and given high effort is called ρ_1.

In order to provide the right incentives, the borrower needs to retain a minimum incentive share, so that the incremental benefit of providing high effort exceeds the private benefit. The fraction of R retained by the firm thus has to equal at least $\frac{B}{\Delta p}$. This limits the exptected amount the firm can promise to outside investors without violating incentive compatibility to the *pledgeable income* (Holmstrom and Tirole, 1997; Holmström and Tirole, 1998) in case of continuation, which equals

$$\rho_0 := p_H \left(R - \frac{B}{\Delta p} \right). \tag{1}$$

Saliently $\rho_0 < \rho_1$. There is a wedge between pledgeable and total income, as the firm needs to retain an incentive share.

Contracts At $t = 0$, investors offer contracts to the borrower. Investors entering into a contract at $t = 0$ are called *incumbent* investors. Contracts with incumbent investors consist of three elements:

- an up-front transfer J from the investor to the borrower,
- a maximum liquidity provision $\overline{L}_i$, at $t = 1$. At $t = 1$, the borrower can demand an amount of liquidity L_i up to $\overline{L}_i$. It is assumed that the borrower has no investment or consumption opportunities at $t = 1$ (as in Holmström and Tirole, 1998), so that the borrower only cares whether or not the aggregate amount of money it attracts is large enough to cover the liquidity shock, and

- a repayment function $D_i : \left[0, \overline{L}_i\right] \to \mathbb{R}^+$, that specifies a debt level for each demanded level of liquidity L_i. In any sequentially rational equilibrium, this function needs to be non-decreasing, as otherwise the borrower would have an incentive to withdraw too much liquidity in some states of the world, wasting the unnecessary amount. We take the debt level to be per unit investment, so that the repayment to each investor equals $D_i(L_i)I$.

Contracting frictions there are two main contracting frictions that play an important role. First of all, the borrower can take on additional contracts at $t = 1$, and the contracts at $t = 0$ cannot stipulate any limitations on contracts at $t = 1$, beyond seniority. Concretely, this means that investors cannot renege on the amount of liquidity they promised based on additional contracts that the borrower might take on at $t = 1$. The only assumption is that claims arising from contracts signed at $t = 0$ are senior to claims coming from contracts at $t = 1$.[2] Second of all, different investors signing contracts at $t = 0$ cannot collude on the contracts they sign with the borrower. This leads to a potential *common agency* problem if multiple investors are needed to contract with the borrower.

At $t = 1$, the borrower can attract additional financing from investors. Investors at $t = 1$ are dubbed *entrants*. The entrants can observe the amount of liquidity the firm attracted from incumbent investors, and thus correctly infer the total liquidity need, as well as the total repayment to both investors combined. Entrants are aware that their claims are junior to those of incumbent investors.

Notation The term *allocation* will be used to describe the combination of investment size, whether or not the firm continues in each state of the world, and, in case of continuation, what the repayment is to incumbents and entrants respectively. In order to discuss equilibrium allocations, it is useful to develop some notation for aggregate quantities. Denote by $\overline{\vartheta}_0$ the maximum liquidity shock the borrower can finance from the funds provided by incumbents. Denote by $\overline{\vartheta}$ the largest shock the borrower can finance using liquidity from both incumbent and entrant investors. For every $\vartheta \leq \overline{\vartheta}$, one can now define $D(\vartheta)$, the total debt repayment (per unit invested) that the borrower has to take on in order to finance a shock ϑ. Denote by $D_1(\vartheta)$ the amount she pays to the entrant when the shock equals ϑ, and by $\vartheta_1(\vartheta)$ the corresponding amount of liquidity obtained from the entrant.

Parameter assumptions there is credit rationing in the sense of Holmström and Tirole (1998): for each of the potential realizations of the liquidity shock, the expected pledgeable

[2] As shown later, if this assumption were to be relaxed, the results would only become stronger.

income is not enough to pay back the total expected cost:

$$f\left(\vartheta_l\right)\rho_0 < 1 + f\left(\vartheta_l\right)\vartheta_l \tag{2a}$$

$$F\left(\vartheta_m\right)\rho_0 < 1 + f\left(\vartheta_l\right)\vartheta_l + f\left(\vartheta_m\right)\vartheta_m \tag{2b}$$

$$\rho_0 < 1 + \mathbf{E}\vartheta \tag{2c}$$

This means that the firm will always need to commit at least part of its own assets to receive any financing.

Furthermore, there is at least one value for $\overline{\vartheta} \in \{\vartheta_l, \vartheta_m, \vartheta_h\}$ for which the ex ante expected total surplus is more than the expected cost, i.e. for which

$$F(\overline{\vartheta})\rho_1 \geq 1 + \sum_{\vartheta \leq \overline{\vartheta}} f(\vartheta)\vartheta. \tag{3}$$

This means it is worthwhile to undertake the project for some cut-off in case of high effort. For low effort, the total surplus generated by the project (including the private benefit) does not yield enough to recoup the cost investment for any cutoff, i.e. for all $\overline{\vartheta}$

$$F(\overline{\vartheta})\left(p_L R + B\right) < 1 + \sum_{\vartheta \leq \overline{\vartheta}} f(\vartheta)\vartheta. \tag{4}$$

It will be useful to define the *effective expected cost of investment* for any cutoff to be the ratio between the total cost of the project (per united invested) and the probability of continuation:

$$c\left(\overline{\vartheta}\right) := \frac{1 + \sum_{\vartheta \leq \overline{\vartheta}} f(\vartheta)\vartheta}{F(\overline{\vartheta})}. \tag{5}$$

As a shorthand, for $s \in \{l, m, h\}$, c_s will sometimes be used instead of $c(\vartheta_s)$. By assumption, this effective cost is minimized for $\overline{\vartheta} = \vartheta_m$, which allows summarizing the conditions above as

$$\max\{\rho_0, p_L R + B\} < c_m \leq \rho_1. \tag{6}$$

As a final assumption, the difference between two subsequent levels of the liquidity shock is small enough, in the sense that both $\vartheta_m - \vartheta_l$ and $\vartheta_h - \vartheta_m$ are smaller than $p_L \frac{B}{\Delta p}$.[3] This means that the borrower can sell his incentive share in order to obtain enough liquidity to bridge the liquidity need between two subsequent states of the world, even though selling this incentive share leads to low effort.

[3] It is important to note that parameter constellations exist that satisfy all the conditions specified above. Take for example the following set of parameters. $p_H = 1$, $p_L = \frac{1}{2}$, $R = 4$, and $B = 1\frac{1}{2}$, with ϑ distributed as $\vartheta_l = 2$, $\vartheta_m = 3$, $\vartheta_h = 4$, and $f_l = \frac{1}{2}$ and $f_m = f_h = \frac{1}{4}$. This indeed gives $c(\vartheta_l) = 4$, $c(\vartheta_m) = 3\frac{2}{3}$, and $\vartheta_h = 3\frac{3}{4}$.

4 Benchmark

We first study the second best allocation, which corresponds to the baseline case studied in Holmström and Tirole (1998). In this benchmark, moral hazard and limited liability are still present, but the allocation is optimized given these two frictions. As is shown in Holmström and Tirole (1998), if the contracting frictions mentioned above are not present, the second best can be implemented by optimal contracting, even though ϑ is not observable.

The optimal $\overline{\vartheta}$ is thus the one that minimizes the effective cost of investment, i.e. $\overline{\vartheta} = \vartheta_m$. This means that, with respect to the second best, policies can either be inefficiently generous with liquidity, allowing continuation even when $\vartheta = \vartheta_h$, or liquidating too often. Furthermore, it is optimal to induce high effort in every state of the world, as well as to always use the full pledgeable income, so as to maximize the firms capacity to attract outside financing. This is stated in the following proposition.

Proposition 1. *In the second best $\overline{\vartheta} = \vartheta_m$, and $p_H D(\vartheta_l) = p_H D(\vartheta_m) = \rho_0$.*

As Holmström and Tirole (1998) argue, this allocation can be implemented by investors at $t = 0$ providing the initial investment, together with an irrevocable line of credit up to ϑ_m.

4.1 Third Best

In the second best allocation, $\overline{\vartheta} = \vartheta_m$ and $D(\vartheta_l) = D(\vartheta_m) = R - \frac{B}{\Delta p}$. As is shown in Boxtel, Castiglionesi, and Feriozzi (2013), this cannot be the case in this paper's setup, as is stated in the following proposition.

Proposition 2. *The second best allocation cannot occur in equilibrium.*

The intuition behind this result is as follows: assume the borrower only deals with the incumbent investors, meaning he can always obtain $\vartheta_m I$ in exchange for a repayment of $R - \frac{B}{\Delta p}$ from them. In that case, a deviation is possible akin to Bizer and DeMarzo (1992), but ex post, rather than ex ante. If a shock of ϑ_h hits, the borrower can obtain ϑ_m from the incumbents and ask an entrant investor to provide the shortfall of $\vartheta_h - \vartheta_m$. In exchange for this additional liquidity, he can offer a repayment of up to $\frac{B}{\Delta p}$. The entrants realize that this will lead to lower incentives, but can still break even on the trade, since $p_L \frac{B}{\Delta p} \geq \vartheta_h - \vartheta_m$. The incentive cost is externalized onto the incumbent lenders, who will still need to provide ϑ_m, but now in exchange for a smaller expected repayment. The entrant thus free-rides on the liquidity

If an entrant is active in equilibrium, it must be that she provides liquidity ex post. Since the total repayment is not increasing in the total amount of liquidity provided, either the entrant provides liquidity for free, or the repayment to the incumbent is decreasing in the amount of liquidity provided by the incumbent. The former case would be loss-giving for the entrant, who would then rather stay out. The latter case would not be incentive compatible for the borrower, as he would then always choose to withdraw too much and burn the excess liquidity in at least one state, in order to have a lower repayment.

In order to find the third best allocation, we define an additional constraint based on the borrower's behavior at $t = 1$, i.e. the allocation needs to be compatible with optimal trading between the borrower and entrants.

Definition 1. An allocation is called *entrant-compatible* if, given the contract with incumbent investors, it induces the optimal trade between the borrower and entrants, given the entrants' break-even constraint. The *third best* allocation is the optimal entrant-compatible allocation satisfying the incumbent investors' break-even constraint.

For an allocation to be entrant-compatible, it must either be that the entrant is active in equilibrium, or that the firm has no incentive to seek additional financing from an entrant. The latter can be achieved in two ways. First, the repayment at the maximum liquidity level could be made so large that it does not allow for an entrant to free-ride in the way described above, i.e. that $p_L\left(R - D(\overline{\vartheta}_0)\right)$ is smaller than the difference between $\overline{\vartheta}_0$ and the next possible liquidity shock. As both $\vartheta_h - \vartheta_m$ and $\vartheta_m - \vartheta_l$ are smaller than $p_L\frac{B}{\Delta p}$, this necessarily means that $D(\overline{\vartheta}_0) > R - \frac{B}{\Delta p}$, so that the borrower would provide low effort in this state of the world.

The second way of making sure the entrant is not active in equilibrium, is to insure the borrower against all liquidity shocks, i.e. to set $\overline{\vartheta}_0 = \vartheta_h$. Since the firm can cover liquidity shocks in all states of the world, he never has a reason to seek additional liquidity.

If the entrant is active in equilibrium, she only becomes so ex post and thus needs to be fairly compensated for any marginal unit of liquidity she provides. If the total repayment to both lenders induces low effort, it needs to be that $\vartheta_1(\vartheta) = p_L D_1(\vartheta)$, and if it induces high effort, it must be that $\vartheta_1(\vartheta) = p_H D_1(\vartheta)$. This means that if the entrant is active in equilibrium, and high effort is induced in all states of the world, the repayment to the incumbents must be strictly below the pledgeable income.

Finding the third best allocation boils down to finding the optimal entrant-proof allocation for each potential level of $\overline{\vartheta}$: ϑ_l, ϑ_m, and ϑ_h respectively, and then comparing the surplus across these possible levels. This gives three potential allocations.

Low liquidity: in this allocation, the incumbent lender provides less than ϑ_l in liquidity and does not require the full pledgeable income in return, leaving enough so that the entrant can provide the remaining liquidity up to ϑ_l, so that the total repayment to both investors precisely equals the pledgeable income. The incumbent reduces the liquidity supply to such an extent that in case of a ϑ_m-shock, there is no scope for the entrant to provide enough additional liquidity in case of continuation. In this case, the total surplus is the maximum that can be attained with low liquidity provision, which is

$$\Pi_l = \frac{\rho_1 - c(\vartheta_l)}{c(\vartheta_l) - \rho_0}. \tag{7}$$

The full surplus accrues to the borrower, effort is high in all continuation states, and the borrower can use all of the pledgeable income, maximizing the outside financing capacity.

Middle liquidity: in this case, the incumbent provides some liquidity, and leaves enough pledgeable income, on the one hand, for the entrant to provide the remaining liquidity up to ϑ_m in an incentive compatible manner. On the other hand, the incumbent needs to make sure that she does not provide so much liquidity that the entrant could provide additional liquidity up to ϑ_h. This gives two constraints limiting both the amount of liquidity the incumbent can provide and the repayment she can demand. In the middle state, the borrower still pays the full pledgeable income ρ_0 to incumbent and entrant combined. In the low state, however, these constraints entail that effectively the pledgeable income is reduced by an amount

$$\Psi := \min\left\{\frac{p_H}{\Delta p}\left(\frac{p_L B}{\Delta p} - (\vartheta_h - \vartheta_m)\right), \vartheta_m - \vartheta_l\right\},$$

which also represents the compensation that the entrant demands for his additional liquidity provision in the ϑ_m-state. The resulting profit is

$$\tilde{\Pi}_m = \frac{\rho_1 - c(\vartheta_m)}{c(\vartheta_m) - \rho_0 + \frac{f_l}{F_m}\Psi}$$

This is similar to the expression for Π_m : the numerator remains the same, as the total surplus generated per unit of investment remains the same and accrues to the borrower. The additional term in the denominator represents the fact that in the ϑ_l-state, the repayment is reduced by Ψ, reducing the equity multiplier and thus the total size of the investment.

Full insurance: finally, the third best allocation can feature continuation in all states of the world. Conditional on full insurance, it is possible to maximize the surplus per unit invested and the income pledged to outside investors by having a repayment of the full pledgeable income in every state of the world. This is entrant-compatible, as the borrower never has any incentive to obtain costly additional liquidity from an entrant lender. This means that the surplus equals

$$\Pi_h = \frac{\rho_1 - c(\vartheta_h)}{c(\vartheta_h) - \rho_0}.$$

These are the three candidates for the third-best allocation.[4] As the proposition below states, which one actually is the third best is determined by which one yields the highest surplus.

Proposition 3. *If*

$$\Pi_l > \max\{\Pi_m, \Pi_h\}, \tag{8}$$

the third best allocation features $\overline{\vartheta} = \vartheta_l$, and $p_H D(\vartheta_l) = \rho_0$.

If

$$\Pi_m > \max\{\Pi_l, \Pi_h\}, \tag{9}$$

the third best allocation features $\overline{\vartheta} = \vartheta_m$, $D(\vartheta_l) =$, and $p_H D(\vartheta_m) = \rho_0$.

If

$$\Pi_h > \max\{\Pi_l, \Pi_m\}, \tag{10}$$

the third best allocation features $\overline{\vartheta} = \vartheta_h$ and $p_H D(\vartheta) = \rho_0$ for all ϑ.

This means that full insurance prevails if both $c(\vartheta_h) < c(\vartheta_l)$, and the reduction in pledgeable income Ψ is so large that it becomes more attractive to offer inefficiently high insurance.[5]

5 Full Insurance and Intermediation

This section contains the main result of the paper: if the third best features full insurance, and investors' endowments are limited, the third best cannot be implemented without an intermediary. If intermediation is allowed, intermediaries arise endogenously.

Under the full insurance allocation, a *common agency* issue can arise. If investors' individual endowments are not sufficient to cover the liquidity shock as a whole, multiple investors need to finance the borrower together. In the third best allocation, the repayment to investors, as a function of liquidity provided, is constant. Even though investors break even ex ante, they provide each marginal unit of liquidity at zero cost. Each investor would prefer the borrower to obtain liquidity from other investors first. An investor can achieve this by raising her marginal price of liquidity, so that the borrower would rather exhaust all the free liquidity from the other lenders first.

To illustrate this issue, assume the borrower is financed by two investors, say 1 and 2, who both provide half the upfront investment, as well as half the liquidity in each

[4] one might also consider the types of allocations that allow for shirking in one state of the world. These, however are never third-best. Having high effort in the ϑ_l-state and allowing shirking in state ϑ_m is dominated by the allocation that only continues in state ϑ_l. Full insurance with shirking in the ϑ_h state is worse than full insurance with work in all three states.

[5] Again, it is important to know whether parameter constellations exist satisfying all of the previous assumptions, that also give full insurance as a third best outcome. Take for example ϑ equal to 3, 4 or 5, with respective probabilities $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{4}$. Again, take $p_H = 1$, $p_L = \frac{1}{2}$ and $B = 1\frac{1}{2}$.

state of the world, and obtain half the pledgeable income in each state of the world. Both investors charge a constant repayment of $\frac{1}{2}\left(R-\frac{B}{\Delta p}\right)$. In this case, investor 1, for example can adjust her pricing in the following way: for any liquidity provision up to $\vartheta_l - \frac{1}{2}\vartheta_h$, she charges $\frac{1}{2}\left(R-\frac{B}{\Delta p}\right)-\varepsilon$ for some small ε. With this contract, once the ϑ_l-shock hits, the firm would rather obtain $\frac{1}{2}\vartheta_h$ from investor 2 and $\vartheta_l - \frac{1}{2}\vartheta_h$ from investor 1. The borrower is still able to obtain the same amounts of liquidity in each state of the world, and pays less in the ϑ_l-state. For investor 1, in state ϑ_l, he provides $\vartheta_l - \frac{1}{2}\vartheta_h$, which is less than $\frac{1}{2}\vartheta_l$. For ε small enough, this means this investor is strictly better off.

This type of deviation is possible as long as marginal liquidity is sold for free. Each investor wants to give a discount in order to make sure the borrower obtains the free liquidity from other investors first. This can be seen as a converse free-riding problem to the one underlying Proposition 2: rather than providing additional liquidity in case of higher liquidity shocks, investors have an incentive to provide less liquidity in case of lower liquidity shocks.

This type of behaviour makes it impossible for multiple investors to provide the third best allocation through multiple bilateral contracting. This result is stated as a proposition.

Proposition 4. *If condition* (10) *holds and*

$$W < \vartheta_h \frac{A}{1+\mathbf{E}\vartheta-\rho_0}, \tag{11}$$

The third-best cannot be implemented through bilateral contracting between the borrower and investors.

5.1 Investors as Intermediaries

If investors are allowed become intermediaries for other investors, an equilibrium with trade not only can be sustained, but arises naturally. Investors endogenously decide to become intermediaries. Assume the third best features full insurance and that endowments are small in the sense of condition (11). If there is no intermediary, then, by Proposition 4, there is either no trade, or there is an entrant-compatible allocation with a surplus below the third best level. In either case, the total surplus going to the entrepreneur is below Π^h. Any investor would have an incentive to become the intermediary. She could do so by collecting money from the endowments of other investors at $t = 0$, which will be repaid with a small return. On the other hand, she can give a contract to the borrower with full insurance and an investment size slightly below the third best level, allowing the intermediary a small profit. As long as the return promised to other investors is not too small, this is a profitable deviation. As long as the difference with the third best investment size is not too large, the borrower's profit can be made close enough to Π_h for this to be attractive to the borrower.

Thus, equilibrium will always feature an intermediary, and this intermediary will always offer full insurance. Moreover, such an equilibrium always exists. This is stated in the following proposition.

Proposition 5. *In equilibrium, one investor becomes an intermediary and offers full insurance to the borrower. The borrower obtains the full surplus of* Π_h*. Such an equilibrium always exists.*

The proof of this proposition follows the reasoning above. The details are in the appendix.

5.2 Risk-Free Savings, Capital, and Investment Risk

This part studies whether or not the intermediated investment as described above can be implemented with savings contracts that are risk-free to the intermediary's depositors. Throughout this subsection it will be assumed that there is no aggregate risk regarding the success or failure of the project, which is equivalent to assuming $p_H = 1$.

In the model so far, there is only a single firm, meaning that liquidity risk is aggregate. In order to offer a risk-free savings account to the depositors, the intermediary must be able to use the funds raised from depositors, combined with her own endowment to first finance investment of I, and then a liquidity shock up to $\vartheta_h I$, and still be able to pay back investors their full deposit. This means that the intermediary must use some of her own endowment as a buffer to potentially finance larger liquidity shocks. The following proposition states the minimum buffer needed.

Proposition 6. *The intermediary can offer risk-free deposits to other investors as long as*

$$W \geq \frac{1 + \vartheta_h - \rho_0}{1 + \mathbf{E}\vartheta - \rho_0} A. \tag{12}$$

The intermediary thus needs to have an endowment that exceeds the cash at hand of the firm. One way to think about this minimum endowment is as bank capital. If the intermediary's endowment is below the limit given in condition 12, the firm could raise equity capital from some investors, and use this as a buffer in order to give risk free deposit contracts to other investors.

Continuing on this, it is interesting to study the amount of capital needed if there is no aggregate risk. Technically speaking, this necessitates a model with multiple firms and a new set of assumptions. Rather than setting up such a model, this is proxied for by keeping the model as it is and assuming that the liquidity need equals $\mathbf{E}\vartheta$ precisely. In this case, we obtain the following result.

Proposition 7. *Absent aggregate risk, the intermediary needs capital of at least* A *in order to offer risk-free savings contracts.*

The intermediary's capital thus needs to match exactly the borrower's inside equity. As the intermediary does not offer any technological improvement to investment, the equity multiplier remains unchanged, meaning the intermediary needs the same amount of inside equity to obtain the third best.

6 Intermediation vs. Other Forms of Financing

This section addresses the trade-off between intermediation and other forms of financing that might arise in this model. There are two different approaches that might shed a light on the direction of causality: it could be that certain technologies call for intermediated financing, or rather that the existence of intermediaries leads to investment being financed differently, and some projects not being financed.

The first approach is to study the equilibria that arise if intermediation is possible. In some cases the equilibrium will feature transaction financing, with financing at $t = 0$, and refinancing at $t = 1$, whereas in other cases the full insurance equilibrium with intermediation will prevail. Finding the parameter conditions for either equilibrium could give an idea as to which technologies give rise to transaction-based financing, and which ones to relationship banking.

The second approach is to explicitly forbid intermediation, and compare the equilibria in this case to the third best ones that are possible with intermediation. It is of particular interest to see which types of projects can potentially not be financed without intermediation that would be financed if intermediaries are allowed.

Any interpretation of parameter conditions has to be done very carefully, as all properties are conditioned on the parameters already satisfying the HT assumptions, as well as on the effective cost of investment being minimized for a liquidity provision up to ϑ_m.

Also, rather than being about the causal direction, the two approaches can be seen as positive and normative, respectively. The first approach states under which parameter conditions we expect banking to occur, whereas the second finds conditions such that allowing intermediation improves surplus.

6.1 Comparing the Third Best Outcomes

As is stated above, the full insurance outcome can only be implemented through intermediaries, as individual endowments are limited. It has also already been discussed above that the other two possible third best outcomes necessarily feature a form of transaction financing: refinancing in case of a shock of $\overline{\vartheta}$ necessarily features ex post dealings with the entrant.

This means that whenever $\max\{\Pi_l, \Pi_m\} > \Pi_h$, it is to be expected that transaction financing prevails, whereas in the opposite case large intermediaries engaging in relationship banking are more likely.

Relationship banking will thus prevail whenever both

$$\frac{\rho_1 - c_h}{c_h - \rho_0} > \frac{\rho_1 - c_l}{c_l - \rho_0}, \tag{13}$$

which is equivalent to stating $c_h < c_l$, and

$$\frac{\rho_1 - c_h}{c_h - \rho_0} > \frac{\rho_1 - c_m}{c_m - \rho_0 + \frac{f_l}{F_m}\Psi}. \tag{14}$$

Condition (13) reduces to

$$\vartheta_m - \vartheta_l + \frac{f_h}{f_m + f_h}\left(\vartheta_h - \vartheta_m\right) < \frac{1}{f_l} \tag{15}$$

whereas condition (14) can be rewritten as

$$f_h\left(\rho_1 - \rho_0\right)\left(\vartheta_h - c_h\right) < f_l\Psi\left(\rho_1 - c_h\right) \tag{16}$$

It can be already be seen from both expressions that whenever the high liquidity shock becomes less probable, the relationship banking equilibrium becomes more likely to prevail. Similarly, if $c_h - c_m$ is very small, these conditions are more likely to hold.

So far, these conditions were considered without taking into account the parameter constraints already imposed when the model was introduced. Taking these constraints, specifically the constraints that $\rho_0 < c_m < \rho_1$, as a given, gives conditions that offer a richer range of interpretations. Therefore, we now study parameter constraints, keeping $\frac{\rho_1}{c_m}$ constant.

6.2 With or without Intermediation

Without intermediation, the full insurance allocation with constant repayment becomes impossible. Instead, there are two potential candidate full insurance allocations that could be implemented through transaction financing: one with reduced pledgeable income in the ϑ_l and ϑ_m-state, and one with full use of the pledgeable income in those states, but with shirking in the ϑ_h-state.

7 Robustness

This section addresses whether the result of the third best featuring full insurance is merely a result of the specific institutional assumptions. As is shown in this section, full insurance becomes only more likely if these assumptions are changed.

7.1 Bankruptcy Arrangements

Above, it was assumed that lenders at $t = 0$ could impose seniority over their claims, meaning that if investors who provided financing at $t = 0$ expect a repayment of D in case of success, the repayment entrants can get in case of success is limited to $R - D$. In order to show the effect of this assumption, the opposite assumption is made here: entrants can arbitrarily dilute incumbents up to the full amount R. It should be noted, however, that dilution can only happen if the borrower promises more than R in total, meaning that dilution necessarily implies low effort. The entrants can expect up to $p_L R$ if they choose to additionally dilute. Note that by assumption $p_L R > p_L \frac{B}{\Delta p}$.[6]

The larger gains for entrants make it possible to extend the liquidity supply under more circumstances. Specifically, this means that potentially the allocations with liquidity provision up to ϑ_m or ϑ_l are no longer entrant-proof. This means that it becomes more likely that the full insurance allocation is the third best.

7.2 Multiple Entrants

Throughout the paper so far, it has been assumed that investors could not control contracts at later dates, but that at the refinancing stage, competition between investors in a classical manner, optimizing the borrower's surplus at that moment. This would be the case if entrants could enforce some kind of exclusivity, or if the refinancing happens in a transparent open market transaction.

In Bizer and DeMarzo (1992), there is an infinite sequence of "entrants" who can all buy any bit of the borrower's stake that is left after the previous investors' debt is paid. In their paper, there is always an inactive investor waiting in the wings, and the allocation needs to satisfy the so-called "no further borrowing" (NFB) constraint that no new entrant is willing to provide any additional financing.

If we take the model from the current paper, but now assume there is an infinite sequence of entrants at $t = 1$, as in Bizer and DeMarzo (1992), then whenever $\overline{\vartheta} < \vartheta_h$, and the repayment $D(\overline{\vartheta})$ to the incumbent and all entrants combined is low enough, there is always scope for an entrant to come in and offer additional financing. Specifically, if $D(\overline{\vartheta}) \leq R - \frac{B}{\Delta p}$, an entrant could always offer additional financing in exchange for a stake worth $p_L \frac{B}{\Delta p}$. This first of all entails that the two allocations with $D(\overline{\vartheta}) = R - \frac{B}{\Delta p}$ and with respectively $\overline{\vartheta} = \vartheta_l$ and $\overline{\vartheta} = \vartheta_m$, are not compatible with the no further borrowing constraint. With full insurance, there is no scope for additional extension of the liquidity supply, so that the full insurance allocation also satisfies this.

Before analyzing which allocation satisfying the NFB constraint is most likely to be

[6] First of all, one needs $R > \frac{B}{\Delta p}$ in order for the pledgeable income to be positive, but it also follows directly from $p_L R + B < p_H R$: subtracting $p_L R$ on both sides gives $B < \Delta p R$.

the "fourth best", it follows immediately that full insurance prevails under a wider set of parameter circumstances, as the two other third best allocation are not possible anymore.

As any allocation with $\overline{\vartheta} < \vartheta_h$ needs to have low effort if $\vartheta = \overline{\vartheta}$, it is impossible to have $\overline{\vartheta} = \vartheta_l$, as that would entail only low effort, with total surplus not exceeding costs. This leaves only one other possible allocation besides full insurance: an incentive compatible repayment in the ϑ_l-state, and shirking in the ϑ_m-state. Conditional on this, it is optimal to maximize the outside financing capacity, giving $D(\vartheta_l) = R - \frac{B}{\Delta p}$, and an additional $\frac{\vartheta_m - \vartheta_l}{p_L}$ paid to the entrant in the ϑ_m-state. This gives a surplus of

$$\Pi_m^{NFB} := \frac{f_l\rho_1 + f_m(p_L R + B) - (1 + f_l\vartheta_l + f_m\vartheta_m)}{1 + f_l\vartheta_l + f_m\vartheta_l - f_l\rho_0 - f_m p_L(R - \frac{B}{\Delta p})} A$$

This gives the following result

Proposition 8. *If $\Pi^h > \Pi_m^{NFB}$, the NFB-constrained allocation features full insurance.*

As $\Pi_m^{NFB} < \Pi_l$, this is less strict than the first best.

Remark on no further borrowing with dilution if one assumes both that entrants can fully dilute previous investors and that there is an infinitely long sequence of entrant investors, only full insurance is possible. Suppose that $\overline{\vartheta} < \vartheta_h$. In that case, an entrant can always extend the liquidity supply by fully diluting all previous investors. This makes any allocation other than the full insurance one impossible under these extreme assumption.

8 Conclusion

A Proofs

Proof of Proposition 1. In the second best allocation, total surplus is maximized. As $p_L R + B < \rho_1$, it is optimal to induce effort in all states in which the firm continues. The surplus to be maximized equals

$$F(\overline{\vartheta})\rho_1 I - \sum_{\vartheta \leq \overline{\vartheta}} f(\vartheta)\vartheta I - I,$$

which equals

$$F(\overline{\vartheta})(\rho_1 - c(\overline{\vartheta}))I,$$

this has to be maximized given the firm's effort constraints: for all $\vartheta \leq \overline{\vartheta}$

$$p_H D(\vartheta) \leq \rho_0,$$

and the firm's break-even constraint

$$\sum_{\vartheta \leq \overline{\vartheta}} f(\vartheta) p_H D(\vartheta) I \geq F(\overline{\vartheta}) c(\overline{\vartheta}) I - A,$$

Note that the break-even constraint should bind, so that

$$I = \frac{A}{F(\overline{\vartheta})c(\overline{\vartheta}) - \sum_{\vartheta \leq \overline{\vartheta}} f(\vartheta) p_H D(\vartheta)}$$

Given that the target function is increasing in I, it is optimal to make sure $D(\cdot)$ is as large as possible in every state of the world, meaning the effort constraint has to be binding. This means the target function can be written as

$$\frac{F(\overline{\vartheta})(\rho_1 - c(\overline{\vartheta}))}{F(\overline{\vartheta})(c(\overline{\vartheta}) - \rho_0)} A.$$

Meaning that the optimal allocation minimizes $c(\overline{\vartheta})$, which occurs for $\overline{\vartheta} = \vartheta_m$. QED

B Some Numerical Examples

References

Becht, M., and C. D. Ramírez (2003): "Does bank affiliation mitigate liquidity constraints? Evidence from Germany's universal banks in the pre-World War I period," *Southern Economic Journal*, pp. 254–272. 5

Bizer, D., and P. DeMarzo (1992): "Sequential banking," *Journal of Political Economy*, pp. 41–61. 4, 10, 18

Boot, A. W. (1999): "European lessons on consolidation in banking," *Journal of Banking & Finance*, 23(2), 609–613. 3

Boxtel, A. v., F. Castiglionesi, and F. Feriozzi (2013): "Non-exclusive Credit Market Competition and Firm Liquidity," Working paper. 2, 10

Calomiris, C. (1995): "The costs of rejecting universal banking: American finance in the German mirror, 1870-1914," in *Coordination and information: Historical perspectives on the organization of enterprise*, pp. 257–322. University of Chicago Press. 5

Calomiris, C., and J. Karceski (2000): "Is the bank merger wave of the 1990s efficient? Lessons from nine case studies," in *Mergers and productivity*, pp. 93–178. University of Chicago Press. 3

Calomiris, C. W. (1993): "Corporate-finance benefits from universal banking: Germany and the United States, 1870-1914," Discussion paper, National bureau of economic research. 5

——— (1999): "Gauging the efficiency of bank consolidation during a merger wave," *Journal of Banking & Finance*, 23(2), 615–621. 3

Chandler, A. D. (1954): “Patterns of American Railroad Finance, 183050,” *Business History Review*, 28, 248–263. 4

Davis, L. E. (1963): “Capital immobilities and finance capitalism: A study of economic evolution in the United States 1820-1920,” *Explorations in entrepreneurial history : EEH*, 1(1). 5

DeLong, J. B. (1991): “Did JP Morgans Men Add Value? An Economist’s Perspective on Financial Capitalism,” in *Inside the business enterprise: Historical perspectives on the use of information*, pp. 205–250. University of Chicago Press, 1991. 4, 5

Dewatripont, M., and E. Maskin (1995): “Credit and efficiency in centralized and decentralized economies,” *The Review of Economic Studies*, 62(4), 541–555. 6

Diamond, D. (1984): “Financial intermediation and delegated monitoring,” *The Review of Economic Studies*, 51(3), 393–414. 5

Gerschenkron, A., et al. (1962): “Economic backwardness in historical perspective.,” *Economic backwardness in historical perspective.* 4, 5

Guinnane, T. W. (2002): “Delegated monitors, large and small: Germany’s banking system, 1800-1914,” *Journal of economic Literature*, pp. 73–124. 5

Holmstrom, B., and J. Tirole (1997): “Financial intermediation, loanable funds, and the real sector,” *The Quarterly Journal of Economics*, 112(3), 663. 7

Holmström, B., and J. Tirole (1998): “Private and Public Supply of Liquidity,” *Journal of Political Economy*, pp. 1–40. 2, 5, 6, 7, 8, 10

Hoshi, T., A. Kashyap, and D. Scharfstein (1991): “Corporate structure, liquidity, and investment: Evidence from Japanese industrial groups,” *The Quarterly Journal of Economics*, 106(1), 33–60. 5

Jeidels, O. (1905): *Das Verhältnis der deutschen Grossbanken zur Industrie: mit besonderer Berücksichtigung der Eisenindustrie*, vol. 24. Duncker & Humblot. 4

Karceski, J., S. Ongena, and D. C. Smith (2005): “The impact of bank consolidation on commercial borrower welfare,” *The Journal of Finance*, 60(4), 2043–2082. 3

Lewis, W. A. (1978): *Growth and Fluctuations 1870-1913 (Routledge Revivals)*. Routledge. 5

Pauly, M. V. (1974): “Overinsurance and public provision of insurance: the roles of moral hazard and adverse selection,” *The Quarterly Journal of Economics*, 88(1), 44–62. 4

Ramirez, C. D. (1995): "Did JP Morgan's men add liquidity? Corporate investment, cash flow, and financial structure at the turn of the twentieth century," *The Journal of Finance*, 50(2), 661–678. 4, 5

Riesser, J. (1910): *Die deutschen Großbanken und ihre Konzentration.* 4