NBER WORKING PAPER SERIES

CUSTOMERS AND INVESTORS:
A FRAMEWORK FOR UNDERSTANDING FINANCIAL INSTITUTIONS

Robert C. Merton
Richard T. Thakor

Working Paper 21258
http://www.nber.org/papers/w21258

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2015

We thank Doug Diamond and Stavros Zenios for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. Robert C. Merton is Resident Scientist at Dimensional Holdings, Inc. He has at various times in the last three years made compensated presentations at meetings hosted by various financial and government institutions.

Customers and Investors: A Framework for Understanding Financial Institutions
Robert C. Merton and Richard T. Thakor
NBER Working Paper No. 21258
June 2015, Revised May 2016
JEL No. D81,D83,G01,G20,G21,G23,G28,H12,H81

ABSTRACT

Financial institutions have both investors and customers. Investors, such as those who invest in stocks and bonds or private/public-sector guarantors of institutions, expect an appropriate risk-adjusted return in exchange for the financing and risk-bearing that they provide. Customers of a financial intermediary, in contrast, provide financing in exchange for a specific set of services, and do not want the fulfillment of these services to be contingent on the credit risk of the intermediary, even when they are not small, uninformed agents lacking in sophistication. This paper develops a framework that defines the roles of customers and investors in intermediaries, and uses the framework to provide an economic foundation for the aversion to intermediary credit risk on the part of its customers. This customer-investor nexus has implications for a host of issues related to how contracts between financial intermediaries and their customers are structured and how risks are shared between them, as well as the consequences of (unanticipated) deviations from the ex ante efficient contractual arrangement for institutional design, regulatory practices, and financial crises. Moreover, customers and investors are often intertwined in practice, and so this intertwining provides insights into the adoption of “too-big-to-fail” policies and bailouts by regulators in general.

Robert C. Merton
MIT Sloan School of Management
100 Main Street, E62-634
Cambridge, MA 02142
and NBER
rmerton@mit.edu

Richard T. Thakor
MIT Sloan School of Management
100 Main Street, E62-615
Cambridge, MA 02142
rthakor@mit.edu

1 Introduction

In three papers, Merton (1989, 1993, 1997) introduced the idea that many types of financial intermediaries provide "credit-sensitive" financial services, i.e., the effective delivery of these services depends on the credit-worthiness of the provider.[1] The intermediary's credit standing can generate externalities for the different business activities of the intermediary, even when they are not directly interconnected through common customers or other such elements. A concrete example is an investment bank that expresses an interest in participating in a bridge loan as an attempt to start a merchant banking business and in doing so runs the risk of having institutional customers flee its over-the-counter derivatives business (e.g. long-dated swap contracts) because of concerns about the ability of the bank to fulfill its contractual obligations on its derivative products were it to suffer losses on its bridge loans (see Merton (1997)).[2] In various theories of financial intermediation, the *raison d'etre* of a financial institution is to serve its customers (depositors and borrowers in the case of a bank, for example), so the potential sensitivity of the perceived value of the intermediary's services to its own credit risk has important implications. Our main goal in this paper is to study how this aspect of the relationship between a financial intermediary and its customers affects the design of contracts between intermediaries and their customers, and how it helps us to better understand commonly-observed features in a wide variety of real-world contracts, institutions, and regulatory practices.

Financial institutions differ from non-financial firms in many significant respects. First,

---

[1] See also Merton (1992a).

[2] A specific example of such a scenario is the case of Salomon Brothers and RJR Nabisco. When Salomon expressed interest in undertaking a leveraged buyout of RJR Nabisco in 1988, many of Salomon's credit-sensitive customers fled because of concerns of how such moves may affect its overall creditworthiness. Salomon's response to mitigate this concern in subsequent years was to create a "ring-fenced" AAA-rated subsidiary called Salomon-Swapco as a counterparty for its OTC customers' derivatives trades.

what investors in a financial institution purchase looks very similar to what its customers purchase. For example, investors who buy a bank's subordinated debt have a financial claim that is similar to the claim of the bank's depositors in that both are debt claims on the bank's cash flows. In contrast, customers and investors in non-financial firms look quite different—a customer who buys a General Motors car is transparently different from an investor who buys the company's bonds. A second distinguishing feature of financial intermediaries is that their assets and liabilities are closely related. For example, in a bank one group of customers—borrowers—provide financing to the bank in the form of deposits (also structured as debt contracts). Thus, the bank has customers with debt contracts on both sides of its balance sheet. The fact that much of a bank's assets and liabilities are comprised of forms of debt also contributes to much higher leverage in banks (and other financial institutions) compared to non-financial firms. In contrast, in a non-financial firm, assets and liabilities typically appear to be very different. In an automobile company, for instance, the assets are manufacturing plant and equipment and its inventory of cars, whereas liabilities are the money the company has borrowed from investors.

A third important distinction is that in financial institutions both investors and customers provide financing to the intermediary.[3] Investors, such as stockholders or bondholders of an intermediary, provide risky financing—they expect the payoffs of their claims to be linked to the intermediary's outcomes. Included among investors are also external guarantors of an intermediary's liabilities. Thus, investors provide financing as well as risk-bearing. Customers, on the other hand, expect services in exchange for the financing they provide, but prefer to not

[3] In non-financial firms, *suppliers* provide the firm with trade credit, which is short-term financing in the form of payables. However, customers end up being *consumers* of finance rather than providers of it. In contrast, in the case of commercial banks, deposits represent customer-financing and make up typically 70%-80% of the bank's total financing.

have the services they receive depend on the fortunes of the provider of the service.[4] Here we need to distinguish between two types of customers. “Credit-sensitive” customers provide financing to the intermediary in exchange for the provision of future services by the intermediary. The financing provided by these customers appears as a liability of the intermediary. These customers derive utility from the services the intermediary provides to them, and this utility is diminished by an increase in the credit risk of the intermediary. Other customers are those who receive financing from the intermediary, such as mortgage or other loan borrowers. They appear on the asset side of the intermediary’s balance sheet, are not credit-sensitive since they have an obligation to repay the intermediary in the future.

Our focus in this paper is on “credit-sensitive” customers (we refer to them as just “customers” henceforth). For these customers, we show that the additional expected return required to induce them to bear the credit risk of the intermediary exceeds that required to induce the investor to bear it. Thus, a financial intermediary that imposes credit risk on its customers will not be able to compete against one that does not. For example, for a whole-life policyholder in a life insurance company to be indifferent to a lowering of the likelihood that the policy will pay off in the event of death, the insurance company will have to increase the expected return on the customer’s investment more than it would have to if it imposed this risk on its investors instead.[5] The key here is not the identity of the economic agent, but the *role* played by that agent, i.e., whether the agent is an investor or a customer who also provides the financing. In some instances,

[4] For example, a life insurance company's policyholders are customers who purchase insurance policies, which provide cash premiums to finance the company’s assets, but also create liabilities for the insurance company. Similarly, depositors in a bank provide (debt) financing for the bank, but they are also consumers of a variety of safe-keeping, liquidity and transaction services.

[5] Survey evidence is in line with our main point. Wakker, Thaler, and Tversky (1997) report that respondents in their surveys said they would pay 20% less for an insurance policy if the probability of default by the insurance company rises from 0% to 1%. Wakker, Thaler, and Tversky (1997) argue that this is hard to reconcile with standard expected utility theory. Our framework provides a rational explanation for such behavior.

the agent may play multiple roles, and may therefore have different expectations of the institution in different roles. For example, a policyholder in an insurance company is a customer but may also hold the company's stock as an investor.[6] This clarifies that the focus of our analysis is *not* on the primitives associated with economic agents—such as their preferences, beliefs, or wealth endowments—but rather what they view as the optimal contract between them and the intermediary in a given role. An efficient contract will need to ensure that the credit risk of the intermediary is borne by the right party, and we argue that this is the investor, *not* the customer.

The questions we address in this paper are: what are the implications of this customer-investor nexus for how the financial intermediaries structure efficient (first-best) contracts with their customers? How does this perspective help us to understand the microfoundations of observed (second-best) contracts between intermediaries and their customers? And does it help us better comprehend certain aspects of financial crises, especially the early triggers?

In our discussion, we make a number of observations that shed light on these questions and also suggest an agenda for future research. First, we begin with a discussion of the first-best contract, and argue that as long as the contract with the customer is creating a positive social surplus, it is optimal for the intermediary to structure the contract so that it is as insensitive as possible to the fortunes of the intermediary itself.[7] This means that the first-best is a contractual arrangement in which the customer is exposed only to the risk inherent in the contract terms—the

[6] As another example, consider an agent who has a deposit account at a bank and also owns equity in the bank, making him/her both an investor and a customer. However, our argument is that the expectation of what the two contracts will provide is different. For the deposit account, the agent cares about the service the bank supplies, and will want to minimize his/her exposure to the credit risk of the bank. In contrast, for the equity share, the agent expects the credit risk of the bank to be reflected in the risk/return tradeoff of the contract.
[7] See also Merton (1997).

risk that the contract itself has specified—and not to the credit risk of the intermediary itself.[8] Put another way, it is not simply that a customer is averse to the *contractually-stipulated* riskiness of cash flows, but that a customer is effectively "risk averse" with respect to the contract itself not being fulfilled. This is something that can occur due to inefficient risk-bearing created by the customer being exposed to the credit risk of the intermediary. At a basic level, exposing the customer to this credit risk is akin to affixing a lottery to the customer's contract which adds unnecessary risk-bearing with no added gain. In contrast to risk-sharing, this added uncertainty does not have social value, as the customer's uncertainty is increased and no other agent's uncertainty is reduced. Investors, by comparison, are in the business of bearing this risk. In a simple analytical example in which customers attach separable values to the monetary flows and the services provided by the intermediary, we explain how exposure of the customer's contract to the idiosyncratic credit risk of the intermediary can result in an inefficient loss of economic surplus. An example of this effect would be a customer that has an annuity contract which is dependent on the fortunes of the issuing firm (and thus is suboptimal for the customer). That customer may optimally want to replicate that contract by purchasing small portions of that annuity from, say, 1000 different intermediaries. However, transactions costs for the customer would make doing so prohibitive. We point out that this desire on the customer's part to not be exposed to the credit risk of the intermediary can also be viewed as a manifestation of a particular form of market incompleteness. We discuss the reasons why these markets are incomplete.

To illustrate this point further, consider briefly some examples; these are developed in more detail in Section 4. A bank depositor is a customer who wants a sure payoff that is not exposed to

[8] In other words, a customer may understand that the contract he/she purchases from an intermediary may have a risky payoff (for example, a mutual fund account that is indexed to the S&P 500). However, the customer does not want any uncertainty about receiving the promised payoff because the intermediary is insolvent.

the credit risk of the bank; those who guarantee these deposits in some way (e.g. the shareholders and the subordinated debtholders implicitly and the deposit insurer explicitly) are investors who bear the risks that depositors do not.[9] Similarly, when a customer purchases an annuity, the customer is guaranteed certain income, even though the value of the annuity is stochastic. By contrast, a bond sold by the same institution represents a claim held by investors who are exposed to the idiosyncratic risk of the institution. Another example is a homeowner who purchases property insurance from an insurance company. The policyholder does not wish to bear any risk of not receiving the promised payoff in the state of the world in which the home is destroyed and the insurance needs to pay off. Shareholders and those who purchase bonds in the company are the insurance company's investors. Yet another example of customers who are not exposed to the intermediary's credit risk are those who invest in a mutual fund, say Vanguard's S&P 500 fund. They receive efficient risk-adjusted returns in a CAPM sense because the risk of the fund not paying out is remote.

Second, there are many possible reasons why this insensitivity is achieved in the efficient contract by creating a *de facto* firewall between the customer's contract and the intermediary's credit risk, rather than by letting the customer bear some of that risk in exchange for a lower price of the contract. Our reasoning proceeds in two steps. In the first step, we argue that the market incompleteness mentioned above constrains the customer's ability to diversify away the intermediary's credit risk if this risk were to be imposed on the customer. As a result, customers are not the best suited to bear this risk—it becomes undiversifiable idiosyncratic risk to customers, and intermediaries that attempt to compensate customers for this risk will not be competitive against intermediaries that instead insulate customers from it. In the second step, we discuss two

---

[9] Note that the intermediary (like a bank) can have customers on both the asset and liability sides of the balance sheet. Out interest is in the customers on the liability side of the intermediary's balance sheet.

reasons why there is this kind of market incompleteness—moral hazard and adverse selection.

The essence of our explanation for the insulation of the customer from the credit risk of the intermediary is that the intermediary is providing the customer a vector of services—a monetary payoff and also a non-monetary service—and this creates a positive economic surplus. Thus, the loss of this surplus due to the failure of the intermediary would represent a net social loss, and the efficient (first-best) contract will be designed to minimize this loss. In the conceptualization of the efficient contract, two pertinent facts about the contract should be noted. One is that the service the intermediary provides the customer involves a *contractual relationship* between the intermediary and the customer, and—for reasons of moral hazard and adverse selection—it cannot be replicated by a claim that is traded in an anonymous market in which its value is independent of the institution offering the service. The other is that the monetary and service components of the contract are often inseparable in the sense that the service is indeed the provision of a specific, contractually-stipulated monetary payoff stream. The combination of these two facts means that the market for the customer's contract is essentially incomplete—one cannot buy a portfolio of primitive claims from *anonymous* agents to achieve the same vector of services provided by the intermediary.[10] Moreover, the economic essence for the existence of the intermediary implies that it is inefficient for the customer to replace the contract with a large number of smaller contract purchased from different intermediaries. This incompleteness necessitates a linkage between the intermediary and the customer's contract and therefore creates a *potential* (off-the-equilibrium-path) exposure to the credit risk of the intermediary that the customer cannot diversify away. Asking the customer to bear this idiosyncratic risk is therefore inefficient and the intermediary attempts to make the customer's contract insensitive to the credit risk of the intermediary.

[10] Indeed, the very existence of the intermediary implies that the market is incomplete, *sans that* intermediary, for its customers.

Third, we turn to the second-best contract, which is constrained-efficient in the sense that the intermediary may face financing frictions—due to adverse selection and/or moral hazard—and thus may find it infeasible or prohibitively expensive to completely insulate the customer from the credit risk of the intermediary. In this case, there is a tradeoff between the loss of efficiency (relative to the first-best) from exposing the customer to the intermediary's credit risk on the one hand, and the cost of insulating the customer from this credit risk on the other hand. Consequently, the second-best contract may expose the customer to some of the credit risk of the intermediary, absent government intervention. Indeed, government intervention in some cases may be rationalized by the desire to eliminate or reduce these costs.

Fourth, we argue that the efficiency of insulating the customer from the credit risk of the intermediary and the possible lack of perfect insulation in the second-best can explain a variety of observed real-world contracts, institutions, and practices. The contracts that are rationalized by the framework developed in this paper are: government-guaranteed (riskless) bank deposits, mutual funds, insurance contracts, and repos in shadow banking.[11] We argue that our framework also produces a new perspective on another financial contract: contingent-capital (CoCo) bonds in banking that have been recently proposed as a way to infuse more risk-absorbing capital into banks during periods of stress (see Flannery (2013), for example).[12] An institution we analyze with our framework is a futures exchange.

Our framework offers insights into some regulatory practices in banking. One of these is the practice of protecting the largest banks in an economy by considering them as "too big to fail" (TBTF) while small banks are allowed to fail.[13] To understand this, we argue that bigger banks

[11] For mutual funds, the exception is if the mutual fund is providing liquidity services for cash.
[12] Adoption in Europe of CoCos is going on now.
[13] See Kaufman (2013) for a discussion of too-big-to-fail.

are more complex than smaller banks in many ways. The complexity of interest to us involves the greater *intertwining* of customers and investors in bigger banks. The bigger the bank, the more difficult it becomes to keep separate investors from customers, so the second-best contract between the bank and its customers exposes customers to more credit risk than in less complex organizations. If the bank is allowed to fail, it also hurts customers as this intertwining means that there is no easy way to protect customers while exposing investors to credit risk. It thus may be necessary to provide guarantees to investors in order to protect customers. The greater loss in net social surplus from the damage sustained by customers in large banks compared to small banks provides a novel economic rationale for too-big-to-fail policies in an environment in which banks rely on second-best contracts with customers and limited government resources force the government to choose to bail out only a subset of banks.[14] This desire to protect customers from the credit risks of their institutions has motivated many government-assisted rescues. For example, the Federal Reserve stepped in to assist Bear Stearns when it was on the verge of collapse in 2008 because of a concern about the possible effect of its failure on the customers of the investment bank; it was involved in trillions of dollars of repo agreements and swap contracts. In the process, of course, the investors of Bear Stearns were also protected. While such interventions are most common in financial services, they have occasionally occurred even for non-financial firms, but for the same logic—the protection of customers.[15] Thus, an important empirical implication of our analysis is that government bailouts are most likely to occur when the failure of the institution is expected to significantly hurt customers and it is difficult to protect customers without also

[14] We want to emphasize that what is at issue here is not just the sheer size of the bank's balance sheet, but rather its complexity that leads to intertwining of credit-sensitive customers and investors and thereby the *systemic* nature of the risk involved in large bank failures. We elaborate on this in Section 5.

[15] For example, the government's bailout of GM and Chrysler during the 2007-2008 financial crisis was also partly motivated by the desire to protect customers.

protecting investors.

Another regulatory initiative we analyze with our framework is the Dodd-Frank Act, a major financial services regulation enacted in 2010 in response to the 2008-2009 global financial crisis. The element of this regulation we focus on is the requirement for swaps to be traded through clearing houses and exchanges. We argue that this feature finds economic foundation in our analysis.

The final part of our analysis extends the framework to examine financial crises, and particularly how they are propagated. A *crisis* can be thought of as an event that is outside of agents' "model of the world" —when the event happens (to which the agents may have assigned an arbitrarily low probability, effectively zero, ex ante), agents cannot comprehend it and are therefore unable to engage in the usual Bayesian belief revision contingent on observing the event. Consequently, they do not quite know how to react. This is somewhat similar to the idea of "neglected risks" in Gennaioli, Shleifer, and Vishny (2012, 2015),[16] and is closely related to the idea of a customer who purchases a contract with an embedded promise of no exposure to the credit risk of the selling intermediary—the customer believes that the contract is first-best and so one does not need to worry about the solvency of the intermediary. When the customer subsequently experiences financial distress of the issuing intermediary, an event that was believed could not happen, the customer has no contingency—even a mental one—for this event. Hence, with the discovery that this is a second-best contract whose value depends on the financial health of the selling intermediary, the customer does not know how to react. This connection allows us to understand how crises may spread through the customer-investor nexus that we describe. To illustrate, suppose that there is a customer who assumes a particular "model of the world" (i.e. the

[16] The psychological foundations for why agents may not fully take into account all their information are provided by Bordalo, Gennaioli, and Shleifer (2012) and Gennaioli and Shleifer (2010).

customer believes that a contract is not dependent on the intermediary's credit risk) but who then observes an outcome that is contrary to his prior belief about the model of the world. This may cause the customer to question his own "first level" prior beliefs about his model of the world, rather than simply arriving smoothly at a Bayesian posterior belief based on the observed outcome. A shift in the first level of priors over the correct model of the world can cause customers to terminate their contractual relationships with the intermediaries, causing a drying up of funding from that source for intermediaries, and triggering/propagating a crisis.

The rest of this paper is structured as follows. Section 2 briefly reviews the related literature. In Section 3, we present the basic framework of a financial intermediary with investors and customers and discuss the first-best contract. We provide various examples involving real-world intermediaries that correspond to our abstract construct. We discuss the economic argument for insulating the claims of customers from the fortunes of the intermediary in the first-best. We then proceed to discuss, in Section 4, why such separation between the contract and the credit risk of the intermediary can be less than perfect in the second-best contract, and explore the steps intermediaries can take to provide customers as much insulation as possible, even in these cases. We show how the degree of exposure of the customer to the credit risk of the intermediary in the second-best contract is determined. Section 5 turns to regulatory practices, and uses the framework to discuss the Dodd-Frank Act and why too-big-to-fail policies make sense even without concerns about systemic risk. Section 6 discusses financial crisis triggers in this framework, with a more formal framework presented in the Appendix. Section 7 provides concluding remarks.

2 Related Literature

Our paper is related to a strand of the literature which rationalizes why bank deposit contracts are designed to be information-insensitive. Gorton and Pennacchi (1990) first explained that agents who lack the skills to efficiently acquire and process information would prefer to invest in instruments that are informationally insensitive, and that depositors were agents of this type. The reason is that when an uninformed agent trades against an informed agent, the uninformed agent loses on average. Therefore, informationally-sensitive contracts are less liquid than informationally-insensitive contracts, and liquidity-seeking agents prefer to invest in deposits that are riskless and hence informationally-insensitive. Since then, others have rationalized debt contracts that are optimally informationally-insensitive for a variety of reasons, including optimal risk sharing. See Dang, Gorton, Holmstrom, and Ordonez (2014), and Holmstrom (2015).

For example, the Dang, Gorton, Holmstrom, and Ordonez (2014) paper relies on the Hirshleifer (1971) notion that information may sometimes not be released because its release can distort risk sharing. In their model, there are two generations of depositors, who are effectively risk-averse.[17] The early generation of depositors will wish to sell their claims to the late generation if hit by a liquidity shock. The liquidity needs make these depositors risk averse to changes in the values of their claims, and so they do not want the late depositors to produce costly information about the value of the bank's assets, since this would make their exit price contingent on this information and hence stochastic. The bank will oblige by not releasing the information it has about its own assets and by investing in opaque assets that would discourage the second generation of depositors from producing information. Thus, the bank in this setting should optimally be opaque—it should withhold information about the bank's risk from depositors. Gorton and Ordonez (2014) use this idea of the demand for information-insensitive debt to argue that a financial crisis can arise if a shock causes holders of such debt to wish to produce information

[17] The assumption is that their utilities are piecewise linear, but globally concave.

about the collateral backing the debt, as this will lead to a decline in output.

A somewhat different perspective on why bank deposits are (optimally) riskfree, at least asymptotically, is provided by earlier work that rationalized financial intermediaries based on informational frictions. In both Diamond (1984) and Ramakrishnan and Thakor (1984), it is efficient for the intermediary to diversify away the idiosyncratic risks of individual loans (or funded projects), so that even if an individual loan that is monitored/screened by the bank remains (partially) opaque, the bank itself becomes riskfree. The optimality of such an intermediary does not depend on depositor risk aversion, however.[18]

Our focus differs from this previous work in a number of ways. First, in our framework, it is not only bank deposits that should be optimally insulated from the credit risk of the bank—and hence insensitive to whatever information might exist about the bank—but *all* efficient contracts between the financial intermediary and its customers should be insulated. This includes a far bigger set of contracts than bank deposits. Second, in our framework, the efficient claim of the customer need not be riskless—it can be risky, but the risk must be confined to the promised contingent payoffs of the contract itself and *cannot* include the credit risk of the intermediary. Third, we address the important question of *why* all of the credit risk and informational risk should be borne by the intermediary's investors in the first-best case, and not by its customers, and we also use our framework to shed new light on issues like too-big-to-fail and the triggers of financial crises. Finally, our main finding that the value of the customer's claim must be independent of the credit risk of the intermediary in the first-best does not depend on information acquisition by customers

[18] And these papers reach essentially the same conclusion of riskfree deposits as Dang, Holmstrom, Gorton, and Ordonez (2014), but reverse the causality in the argument—the bank is not opaque because it wants to appear informationally-insensitive to its depositors, but rather it diversifies away the idiosyncratic risk associated with each individually-opaque asset it monitors/screens in order to reduce contracting costs and thus asymptotically eliminate risk for its depositors, so that its overall asset portfolio is indeed transparently riskfree to its depositors.

being costly or inherently inimical to stability. Rather, our approach suggests that in well-functioning markets, customers (such as depositors) do not have a *need* for their contracts to be opaque, since their contracts should be optimally structured to insulate them from the risks of the service-providing intermediaries. While opaqueness may be beneficial to producers (e.g. banks), our analysis suggests that it need not be beneficial to customers. In other words, customers will be indifferent to opaqueness as long as their claims do not depend on the fortunes of the intermediary. Therefore, in high-quality debt markets, it need not be the case that transparency causes dysfunction or that opaqueness is necessary.

Our paper is also related to the literature on the functional perspective in finance (for example, Merton (1993, 1995), Merton and Bodie (1995), and Merton and Bodie (2005); see Campbell and Wilson (2014) for a review). In particular, our focus in this paper is on the functions that financial intermediaries serve in meeting the needs of their customers. Thus, while we list specific contracts and institutions as applications of our framework, these serve mainly as examples of the *functions* we seek to highlight in the customer-intermediary interaction. This is in line with the functional perspective, which also focuses on the economic functions served rather than the specifics of the institutions that serve them.

3 Financial Intermediaries and Customers

3.1 Motivation

Financial intermediaries have customers and investors as key stakeholders. As we pointed out in the Introduction, the crucial difference between a customer and an investor lies *not* in who provides financing—both do—but in who bears intermediary-specific risk. In what follows, we will show that the efficient (first-best) contract is such that investors bear this risk, but customers do not. By "first-best", we mean that the intermediary does not face any financial frictions in raising financing from investors or in purchasing guarantees in the capital market that can insulate its customers from its own credit risk. The labels "customer" and "investor" refer to roles rather than individual. Thus, an agent may be *both* a customer and an investor of the same institution, by purchasing products/services and also providing risk-bearing financing. It is, however, essential to keep separate the two roles and not attempt to combine them in cases where the same agent serves both roles. In our discussion, we will generally refer to a customer-financier as simply a "customer".

Customers of an intermediary may want services where the outcome is risky in the sense that the payoff is random, but they do not want the outcome to be dependent on the credit risk of the intermediary. For example, a customer of a brokerage firm who purchases a share in the S&P 500 through that broker expects the performance of that purchase to be uncertain. However, the customer does not want the position to be dependent on the fortunes of the brokerage firm. Customers who have this risk are transformed into investors, which they do not wish to be.

As discussed in the Introduction, in a financial intermediary, a distinction can be made between two types of customers: "credit-insensitive" customers—those whose interaction with the intermediary is such that they are largely unconcerned about the intermediary's credit risk—and "credit sensitive" customers, who care about the intermediary's credit risk. Credit-insensitive customers do not provide financing to the intermediary and essentially pay for what they purchase after they purchase it, which means the intermediary's credit risk does not have an impact on the supply of the service to them by the intermediary. Examples include an individual who asks a broker to purchase a security on his behalf and pays the brokerage commission when the trade is executed, a home seller who pays a commission to the real estate broker (intermediary) when the house is sold, and a borrower who gets a bank loan. Credit-sensitive customers are those who provide the intermediary financing by essentially paying for the services up-front, and subsequently receiving these services over time. Because the receipt of the services that have already been paid for is contingent on the survival of the intermediary, the value of the services to the customer becomes dependent on the intermediary's credit risk. Examples are bank depositors, insurance company annuity and life insurance policyholders, and those who purchase standby letters of credit from banks. Thus, both types of customers purchase services from the intermediary—for example, a bank's depositors purchase the service of having a safe place to keep cash that is available on demand and accumulates interest, whereas its borrowers view the bank as a source of loanable funds. Our focus is on credit-sensitive customers, who we refer to as "customers" for short.

It should be noted that the customers we focus on in our analysis are not simply "widows and orphans" or unsophisticated investors. A customer could be a large institution, such as the World Bank or a large pension fund. Thus, the desire to be insulated from the credit risk of the intermediary is not limited only to customers who are informationally-disadvantaged in trading with informed investors, as in Gorton and Pennacchi (1990). The key is that the customer demands a specific set of services, and is averse to any risk that the set of services specified by the intermediary may not be fulfilled. What we argue is that this aversion does not necessarily spring from primitive preferences (i.e. concave utility or risk aversion), but rather is rooted in the efficiency of the intermediary's contracts with its customers and its investors. In essence, when the customer is exposed to the credit risk of the intermediary, an unnecessary lottery is added to the contract, which makes it inefficient for the customer. An intermediary's investors are the ones who are willing to take this risk. So the World Bank may want to hedge its interest rate exposure costs efficiently, and thus hold large exposures to those rates through relatively few financial intermediaries. It can only do so prudently if it is satisfied that the promised payments on its swaps will be made.

3.2 Analytic Setting: Efficient Customer Contracts

To fix ideas more concretely, we now introduce a simple analytic example to help the exposition of our analysis. Let V be the value of the service that an intermediary provides to its customer. It is the monetary equivalent of the expected utility (or the certainty-equivalent of the expected utility) that the customer gets at $t = 0$ from the intermediary's services, and can have many components, as we discuss below. Thus, if the customer is a depositor, then V could represent the monetary value the depositor attaches to having access to a liquid claim at a moment's notice, being able to write checks against the deposit account to conduct transactions, availing of safe-keeping services associated with being able to deposit money in a secure place, etc. For a policyholder in an insurance company, V could represent the value of the utility the individual derives from being able to insure against an accident or a catastrophic event like death. In all of these cases, the contract calls for the customer to provide a set of payments f_t to the intermediary at various dates $t \in [0, T]$, where $[0, T]$ is the period over which the contract exists, in exchange for a vector of services that may include future monetary payments.

More specifically, from the perspective of the customer, V includes two components. The first component is V_m, which is the monetary equivalent of the utility that the customer derives based on the net monetary flows between the customer and the intermediary—i.e., the money f_t flows from the customer to the intermediary, and whatever (possibly state-contingent) money F is paid by the intermediary to the customer as part of the service. In a bank, F could be the amount of deposits (plus interest) withdrawn by the depositor.[19] In an insurance context, F would be the payment made by the insurance company in the event of an accident or death. While F may be deterministic, it can also be stochastic. The second component is V_s, which is the monetary equivalent of the utility the customer derives from the services provided by the intermediary. As an example, if the customer is a bank depositor, then V_m would be the monetary equivalent of the depositor's utility from receiving interest on the deposit (the difference between what the bank returns to the depositor and what was deposited in the bank), whereas V_s would include the monetary equivalent of the utility associated with check-writing privileges (access to liquidity), safe-keeping services, cash management advice, etc. Put together, the two components sum up to V, so $V_m + V_s = V$.

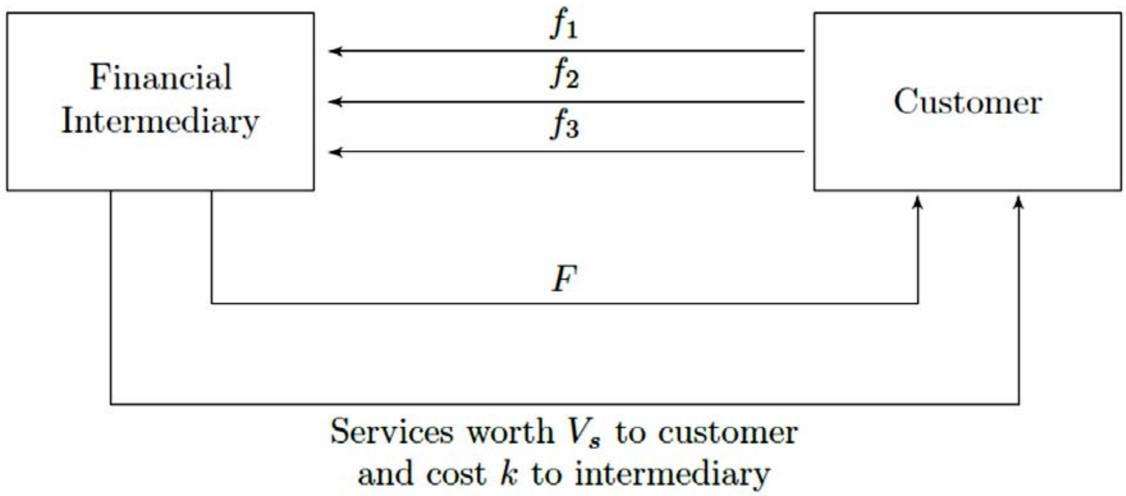
Now define $\overline{V}$ to be the monetary equivalent of the reservation utility of the depositor—it will capture the opportunity cost for the depositor to use bank deposits rather than invest the money elsewhere, say in a money market mutual fund. Satisfaction of the depositor's participation constraint requires

$$V_m + V_s = V \geq \overline{V} \tag{1}$$

Let $k > 0$ be the cost to the intermediary of providing the vector of services that the customer values, and V_m^I the monetary value (in dollars) at $t = 0$ to the intermediary of obtaining financing from the customer. *Figure 1* below describes the relationship between the intermediary and the customer in terms of the values and costs of the financing provided by the intermediary and the value of the services provided by the intermediary.

[19] Put another way, V_m is the monetary equivalent value of the expected utility from $F - \text{PV}(\sum_t f_t)$, which represents the present value of all monetary flows. There need not be only one withdrawal. With multiple withdrawals, F would be the present value of all withdrawals.

Figure 1: Values and Costs of Financing and Services



We assume that

$$V_m^I - k > 0 \tag{2}$$

Taken together, (1) and (2) imply that intermediation creates a positive net economic surplus. This net surplus (in dollars) is

$$V - \overline{V} + V_m^I - k > 0 \tag{3}$$

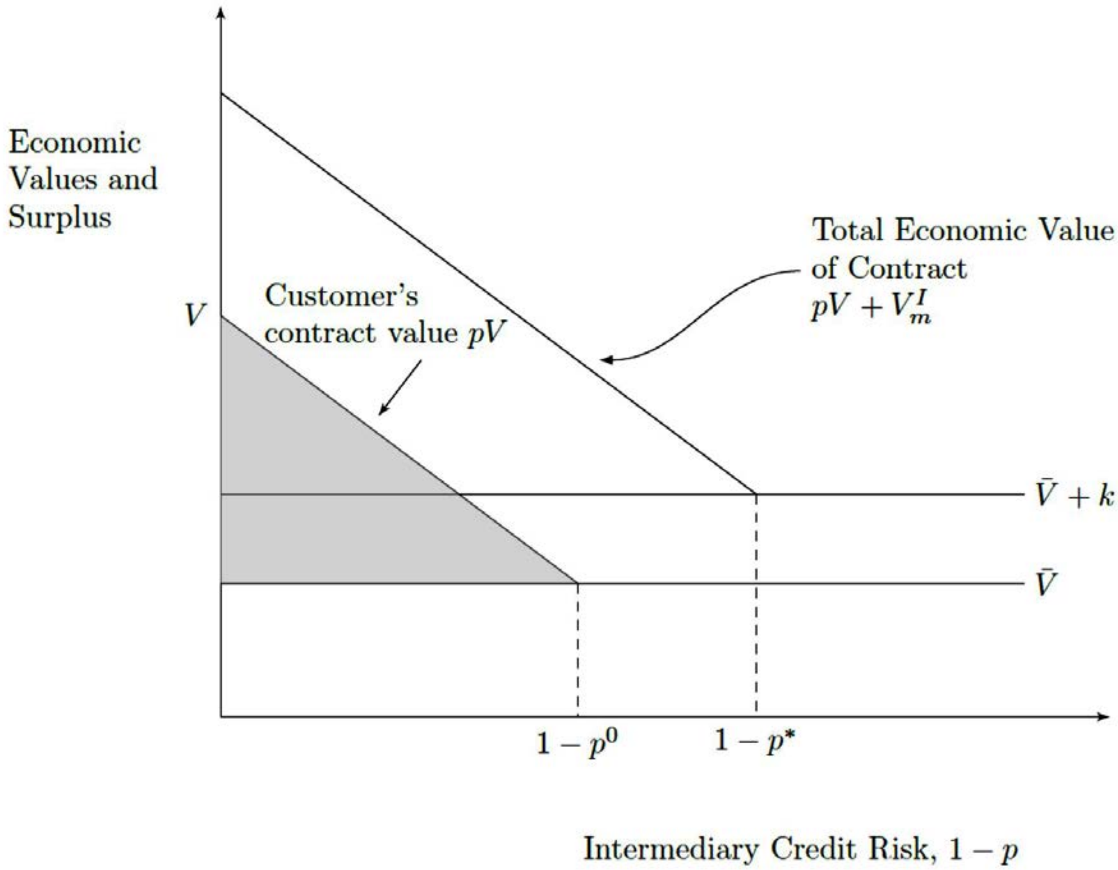
We can attribute this surplus to the specialization-related skills that provide the economic rationale for the existence of the financial intermediary. Let the duration of the contract between the intermediary and the customer be over the time period $[0,T]$.

For simplicity of exposition, suppose the contract is entered into at $t = 0$, at which date the customer provides financing, and then the contract is fulfilled at a single date $t = T$, at which time the intermediary provides all of the services the customer values at V. Let $p \in [0,1]$ be the probability that the intermediary will be solvent at $t = T$, and only if it is solvent can the services valued by the customer be provided. Thus, $1 - p$, the complement of this probability, represents the idiosyncratic credit risk of the intermediary that the contract is exposed to. The value of the contract to the customer now becomes pV, and the participation constraint now becomes $pV \geq \overline{V}$. Thus, the customer's net expected economic surplus relative to its other options is $pV - \overline{V}$. This net surplus is $V - \overline{V}$ if there is no credit risk, which means that the expected loss of net economic surplus due to the intermediary's credit risk is $[1-p]V$. The total expected value (to both the intermediary and the customer) due to the contract is $pV + V_m^I$, and the total net expected economic surplus considering the intermediary's cost of service provision k and the customer's alternative to the contract is $pV + V_m^I - [\overline{V} + k]$.[20] If there is no intermediary credit risk, the net economic surplus is $V + V_m^I - [\overline{V} + k]$. This means that the expected loss of net economic surplus due to the credit risk of the intermediary is $[1-p]V$, a quantity that is increasing in the intermediary's credit risk, $[1-p]$. We call this a "customer contract fulfillment" (CCF) cost. The efficient contract drives this cost down to zero.

Figure 2 below shows how the total economic value of the contract (to the customer and the intermediary) surplus and the customer's share of this total contract value behave as functions of the intermediary's credit risk, $1-p$. It is easy to verify that the total economic value line has the same slope as the customer's contract value line. In *Figure 2*, for any given contract, $1-p^0$ is the intermediary credit risk at which the customer's net expected economic surplus, $pV - \overline{V}$, becomes zero, and $1-p^*$ is the intermediary credit risk at which the total expected net economic surplus $pV + V_m^I - [\overline{V} + k]$ becomes zero. It is straightforward to verify that $1-p^* > 1-p^0$.

[20] V_m^I is not multiplied by p in these expressions because all financing is provided by the customer up front at $t = 0$. Thus, insolvency on the part of the intermediary at a later date will not reduce the value to the intermediary of obtaining financing from the customer.

Figure 2: The Effect of Intermediary Credit Risk on Contract Value and Expected Net Economic Surplus



Two points are worth noting. First, if the intermediary exposes the contract to its own credit risk, the customer cannot recover the entire loss of surplus by hedging this risk—say by buying a put option on the intermediary. The reason is that such risk mitigation can prevent the expected loss of at most $[1-p]V_m$ of the contract value to the customer, as the expected loss of the service portion of the contract value to the customer, $[1-p]V_s$, is unrecoverable. This results in a value wedge or deadweight loss in terms of economic surplus. To ensure that the surplus related to this part of the contract value is not lost, the intermediary has to be solvent at $t = T$.[21]

[21] Thus, another way of thinking about V in relation to the earlier discussion, is that V_m could be viewed as the monetary equivalent of the standard utility over wealth for risk-taking (i.e., the monetary flows that the contract stipulates is

Second, this suggests that the more efficient solution is for the *intermediary* to undertake risk mitigation to insulate the contract from its own credit risk, rather than expect the customer to do it. Merton (1997) identifies various ways in which the intermediary can do this: hedging, purchasing insurance from a guarantor, and keeping sufficient equity capital on its balance sheet.

It is important to note that this result does not depend on risk aversion, in the traditional sense, on the part of the customer. Risk aversion (for example, making the customer infinitely risk averse) may be one particular way to capture this phenomenon. But if one resorts to this explanation, then it should be emphasized that this would be risk aversion *with respect to the uncertainty about the ability of the intermediary to deliver the embedded promise in the contract itself*, and not necessarily the randomness in the final payoffs that contract might specify the customer would be exposed to. For example, a customer may indeed *expect* the final payoffs of the contract to be risky (as in a stock index mutual fund or a swap contract), but it is not risk aversion with respect to these payoffs that should play a special role in any explanation based on the risk aversion of customers. That is, the normal concept of risk aversion from holding stocks and bonds does not accurately capture the behavior of customers that we are discussing here, where we are comparing the efficacy of alternative contracts with the financial intermediary.

3.3 The Inefficiency of Risk Mitigation by the Customer

There are two ways in which a financial intermediary's customer can mitigate the intermediary's credit risk if such risk is imposed on the customer: (i) by diversifying across many intermediaries, or (ii) by accessing an Arrow-Debreu market in primitive state securities to

risky), while V_s can be viewed as a separate component for the services the intermediary provides, which the customer wants to be credit-insensitive.

replicate the vector of services provided by the intermediary without being exposed to the credit risk of the intermediary. We explain now why both are either inefficient or infeasible.

First, consider (i). To diversify away the intermediary's idiosyncratic credit risk, the customer would have to replace the contract that it has with a single intermediary with a large number of smaller contracts with many intermediaries. However, one reason why we have financial intermediaries is that they achieve economies of scale and scope and reduce transaction costs; in our model, this would be reflected in k being, say, invariant to or concave in the size of the intermediary's contract with the customer. Thus, any attempt on the customer's part to diversify across intermediaries will inherently be inefficient due to transaction costs.

Now consider (ii). Our argument is that replicating services is infeasible because of market incompleteness in contracting. The incompleteness is that the customer cannot purchase a separate (Arrow-Debreu) claim that would deliver the service that the intermediary provides when it is solvent. That is, the intermediary is unique in providing its service once it has entered into a contract with the customer. In a complete market, the monetary and service component of the intermediary's contract would be traded separately as bundles of primitive Arrow-Debreu claims. This would enable the customer to purchase market-based insurance against the intermediary's credit risk. But this is often physically impossible because the service the intermediary provides is inseparable from the monetary component of the contract, as explained in the Introduction.[22]

But even if physical separability of these two components was possible, markets for intermediary services to customers would not be complete because the service is something that has to involve a contractual *relationship* between the intermediary and the customer—it cannot be something remote from the intermediary that can be traded in an anonymous market and purchased

[22] The possibility of purchasing such primitive claims (with no contract risk) to replicate the desired payoff would mean that there would be no economic role for the financial intermediary in the first place.

by the customer. This essential coupling of a specific intermediary with a specific customer often generates valuable customer-specific information that is available privately only to the intermediary, information that the intermediary can use to enhance the value of its service to the customer. This rules out a complete market in which state-contingent claims can be created with values that depend only on states of the world and not on the "institutional affiliation" of each claim.[23]

One could argue that one way to get around this problem would be to have a third party purchase the service the intermediary is providing the customer and have it as part of a separate contract that is now decoupled from the intermediary. However, such an arrangement suffers from moral hazard and adverse selection frictions, as we discuss next.

3.4 Moral Hazard as a Source of Market Incompleteness

Another reason why the optimal contract should be insulated from the credit risk of the intermediary is *moral hazard*, which is an important element of many contracts (e.g. Holmstrom (1979)). We argue in this section that it can stand in the way of a complete market in customer contracts. In the framework described above, moral hazard can come in two forms. If k, the intermediary's cost of providing services is unobservable and cannot directly be contracted upon, then the intermediary may underinvest in service provision. For example, a depositor in a bank may be a small business that views cash management advice as part of the services the bank provides depositors, and the bank can underinvest in the personnel capable of providing this advice. Another example is a mutual fund company that offers its customers the choice of

[23] This is somewhat similar to the idea in Froot and Stein (1998) that financial institutions hedge the risk of illiquid assets in the capital market.

allocating their savings across multiple mutual funds managed by the company.[24] Part of the services valued by the customers may be the advice fund employees can offer on how the customer should allocate funds, based on personal financial goals. The fund company can underinvest in the provision of this service by employing fewer people, so customers have to experience long waiting times when they call. The key here is that the service the customer receives has already been paid for by the customer, so the usual contractual resolution of moral hazard, namely conditioning the intermediary's payment on the service provided, does not work here.

A second kind of moral hazard arises from the intermediary's incentives to take excessive risk. Because the customer's financial claim is a variant of a debt contract, the intermediary's (shareholders') payoff is isomorphic to a call option on the total payoff of the intermediary. This can lead to excessive risk taking at the expense of the customer.[25]

While insulating the contract from the intermediary's credit risk will not fully resolve the first kind of moral hazard, *not* insulating the contract will worsen it. The reason is that a financially-distressed intermediary has stronger incentives' to underinvest in serving the customer.[26] It is clear that insulating the contract from the credit risk of the intermediary completely resolves the second kind of moral hazard.

Moral hazard thus provides another powerful reason for the intermediary to pre-commit to *not* expose its customers to its own credit risk, and take steps to either mitigate it in the financial market or transfer it to the shareholders (pure investors). Doing this not only lessens the moral hazard but also preserves more of the surplus the intermediary is able to create by contracting with

---

[24] For example, Fidelity and Vanguard give their customers the ability to split their investments across dozens of funds.
[25] See Merton (1977), but this may not always be the case (e.g. Merton (1992b)).
[26] While the ship is sinking, the intermediary's attention is likely to be focused on cost-cutting rather than customer service, since this conserves financial resources.

the customer, thereby reducing the customer contract fulfillment cost. Of course, in the second-best case, moral hazard may impede the ability of the intermediary to insulate its customers from its own credit risk, and may thus distort the outcome away from the efficient separation between the intermediary-customer contract and the intermediary's credit risk achieved in the first-best

3.5 Adverse Selection as a Source of Market Incompleteness

While different intermediaries may offer exactly the same contract to customers, so that customers can evaluate the contract independently of which intermediary is offering it, the intermediaries themselves may be quite different from each other. As long as the contract is insulated from the credit risk of the intermediary, this cross-sectional heterogeneity is irrelevant. But if the credit risk of the intermediary is commingled with the risk embedded in the contract payoffs to the customer, then the customer needs to discover the credit risk of each intermediary. This may be costly, so not insulating the contract from the credit risk of the intermediary can result in wasteful investigation costs.

But discovering each intermediary's credit risk may not be enough. The customer may need to also learn *how* this credit risk will affect the contract between the intermediary and the customer. This may be very difficult, so in the end it may prove to be prohibitively costly for the customer to acquire the necessary information about the true risk in the contract through its exposure to the credit risk of the intermediary. In such a situation we will have *adverse selection*—each intermediary knows more about its own credit risk and the implications of this for its contract with the customer than the customer does.

As Akerlof (1970) showed, adverse selection can cause a market breakdown. In our framework, suppose that it is common knowledge among potential customers that the value of p

varies in an *a priori* unobservable way among intermediaries. Each intermediary knows its own p privately. Others only know that p is distributed in the cross-section according to a probability density function $h(\cdot)$. Then, if all financial intermediaries offer the same (pooling) contract to customers, they will value it at

$$\int Vpf(p)\, dp = \overline{p}V \tag{4}$$

And for $\overline{p} < p^*$, no contracting will occur because the total economic surplus will be negative. In such instances, financial intermediaries will have a strong incentive to insulate their contracts with customers from their own credit risk.

This insulation, if it can be achieved, will succeed in allowing intermediaries to raise financing from customers. But what will be the impact of adverse selection on the intermediary's ability to raise financing from *investors*? In other words, the more the intermediary insulates the customer from adverse selection, the more of the adverse selection it shifts to the investors. This has not been an issue thus far because we have been discussing the first-best contract where the intermediary does not face any frictions in raising financing from investors. However, in the second-best contract, when such frictions are encountered, an important question arises: *why is it economically more efficient for the risk/cost of adverse selection to be borne by investors rather than by customers*?

This question is central to understanding why the efficient contract involves intermediaries insulating customers from their own credit risks, and what the limits of this insulation might be in the second-best case. We know from the work of Myers and Majluf (1984) that adverse selection in the market for raising financing can generate serious distortions, including firms eschewing positive-NPV projects. So, why are financial intermediaries willing to expose themselves to these potential distortions in order to insure customers against this risk?

An economic rationale for this lies in the fact that securities are *traded* in the capital market, whereas the claims of customers are *not*. We explained earlier, in the context of market incompleteness of the contract between the intermediary and its customers, why the claims of customers are typically not traded. Right now, we focus on how this tradability distinction between investors and customers leads to the optimal allocation of risk between them.

When the financial intermediary raises financing from investors, the adverse selection creates incentives for some agents to invest in acquiring privately-costly information to learn which intermediaries in the pool of *ex ante* observationally identical intermediaries are more highly valued. In a market microstructure model such as Kyle (1985), the informed agents are able to earn trading profits due to their knowledge of which firms are overvalued and which are undervalued, and this is possible because noise trading prevents all of the information possessed by the informed traders from being reflected in market prices.[27] This preserves the incentives for some investors to become informed, and also results in some of the adverse selection problem being dissipated.[28] This is because although not all of the information of the informed investors is reflected in prices, *some* of this information is incorporated into prices due to rational inferences by the market maker based on observing the total order flow in the equity of the intermediary. That is, due to the trades of informed investors, intermediaries with different (privately-known) values trade at different prices in equilibrium, with higher-valued firms trading on average at higher prices. Consequently, active trading helps to lower the adverse-selection costs that intermediaries face in raising capital from investors.

---

[27] We could think of noise traders whose trading motives are not related to private information about asset values and are thus exogenous and random.

[28] It is also possible that each informed trader has a *different* piece of information, so that the market-clearing price aggregates over all these disparate pieces of information and the market thus "knows" more than any one individual, or even the intermediary itself. This is the essence of Hayek's (1945) case for the superiority of decentralized market outcomes over central planning.

The contracts between customers and intermediaries are typically not traded. A simple reason for this is that different customers may assign different values to the services they consume, and these values may be privately known. That is, V_s may vary in the cross-section of customers in a way that is not publicly observable.[29] This will impede the tradability of these contracts. Lack of trading means that adverse selection cannot be diminished for these contracts the way it can be for capital raised from investors.

The fact that investors' claims are traded also means that they can diversify away the idiosyncratic risk of the intermediary. In an efficient market, each risk should be borne by those with the greatest risk-bearing capacity for that risk. Since customers cannot diversify (or hedge) the intermediary's idiosyncratic credit risk but investors can, it is investors who should bear this risk in an efficient arrangement.

This discussion indicates that even the second-best contract will seek to protect customers against the credit risk of the intermediary, because customers will continue to display extreme aversion to being exposed to this risk. Nonetheless, we cannot rule out the possibility that financing frictions faced by intermediaries may be so large that the second-best contract may involve the customer bearing *some* of the credit risk associated with the intermediary, i.e., the second-best contract may involve a positive CCF cost, as we discuss in the next section.

3.6 The Constrained-Efficient (Second-Best) Contract

In the analysis in the previous sections, it was implicitly assumed that the cost to the financial intermediary of insulating its customers from its own credit risk—via hedging,

[29] Another possible reason is that the contract is intended to provide risk-sharing services to the customer, and trading of the contract may destroy optimal risk sharing. For models along these lines, see Diamond and Dybvig (1983) and Dang, Gorton, Holmstrom, and Ordonez (2014).

purchasing insurance from a guarantor, or keeping sufficient equity capital—was less than the benefit of the insulation to the customer. But there may be a variety of costs the intermediary faces in providing its customers complete insulation from its own credit risk. We discuss these below.

3.6.1 Equity Capital

One way that the intermediary can protect its customers from its own credit risk is by putting equity on its balance sheet. With sufficient equity, the intermediary can minimize its credit risk. However, as has been well-established in the literature, the intermediary may face costs with respect to keeping a significant amount of equity on its balance sheet.

Adverse selection costs are one reason why it may be costly for the intermediary to raise external equity (as explained by Myers and Majluf (1984)). Moral hazard may be another reason why the intermediary may find it costly to keep a lot of equity on its balance sheet. Hart (1995) explained how debt can discipline the firm by imposing a periodic fixed payment obligation that provides incentives for the firm to generate cash flows that are sufficient to meet debt repayment. Equity does not have this feature, so by using substantial equity, the intermediary may forgo this discipline on its managers. There may also be other costs of keeping equity. The tax shield associated with debt is one widely discussed reason, for example. Moreover, for intermediaries like banks, implicit bailout protection offered by the government can make debt financing "artificially" cheap compared to equity (which lacks such protection).

All of these costs of equity relative to debt can mean that the amount of equity needed to trivialize the intermediary's credit risk may be so large that it causes the intermediary's capital structure to deviate too far away from its private optimum to make it optimal for the intermediary to keep that much capital.

3.6.2 Hedging and Purchasing Insurance from a Guarantor

If there is moral hazard in the sense that the intermediary can choose (privately costly) hidden action that can affect the returns to investors, then investors will reflect that in the pricing of securities.[30] This pricing is a source of capital market discipline, and it can move the intermediary closer to choosing the action that maximizes the (financial) value of the intermediary. Examples of such discipline include the actions creditors can take, such as monitoring and restrictions imposed on the intermediary when covenants are violated.

However, this discipline imposed by investors can be supplemented with discipline imposed by the intermediary's *customers*. We argued in the previous sections that customers will display extreme aversion to being exposed to the intermediary's credit risk. This aversion too can be a source of market discipline if customers are even partially exposed to the intermediary's credit risk. That is, if the intermediary does not devote enough resources to significantly reducing its credit risk and customers are partially exposed to it, they will flee the intermediary, as discussed in the Introduction (see Merton (1997)).[31] This threat of loss of customers can provide additional discipline on the intermediary. By purchasing insurance from a guarantor or by hedging, the intermediary could choose to completely insulate the customer from its own credit risk, but then market discipline from customers would be lost. So the second-best contract may involve customers being (at least potentially) exposed to some credit risk.

3.6.3 Customer Risk Exposure in the Second-Best Contract

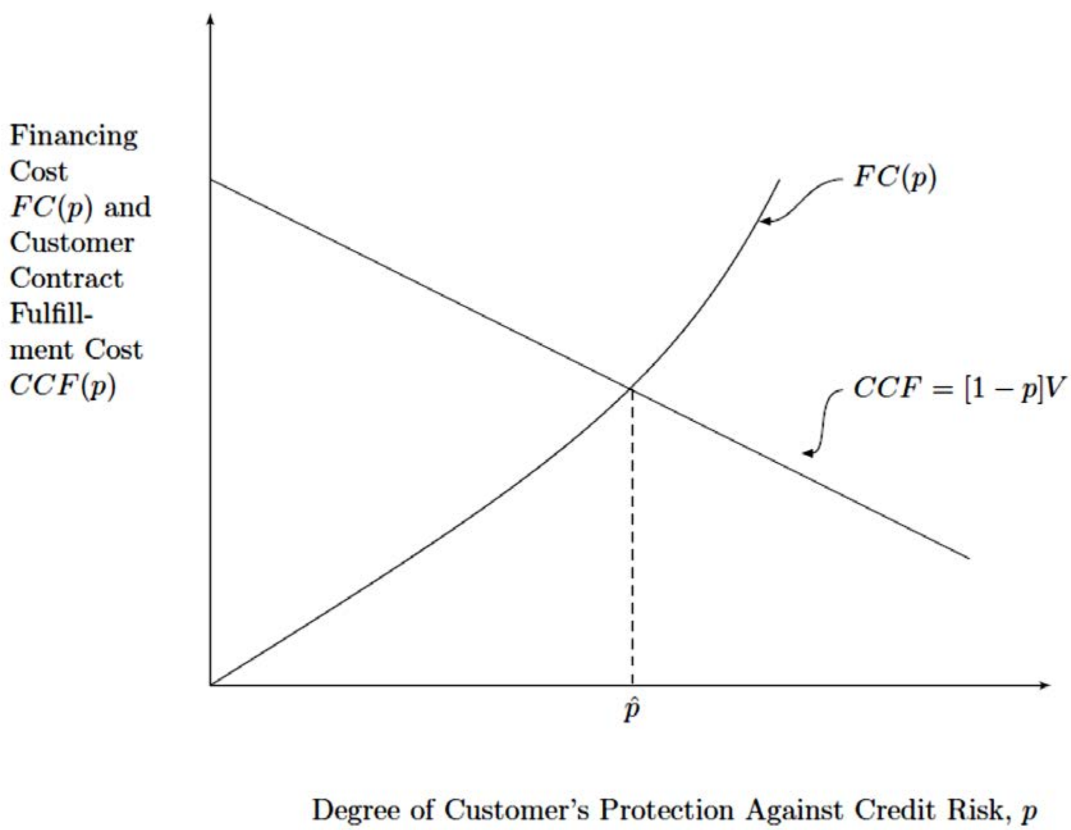
In more analytic terms, let $FC(p)$ represent the "financing cost" faced by the intermediary

[30] Examples of such hidden action are the effort choices of the intermediary's managers, the risk profiles of the projects the intermediary invests in, the resources it devotes to risk management, etc.

[31] This includes instances in which this credit risk exposure results from inefficient choices by the bank. Calomiris and Kahn (1991) model such a situation. In their model, (uninsured) depositors flee the bank if they observe/suspect that the bank manager is making bad investment decisions, and this threat disciplines the manager.

in reducing the customer's exposure to the credit risk of the intermediary, where $1-p$ is the customer's credit risk exposure. We assume $\partial FC / \partial p > 0$ and $\partial^2 FC / \partial p^2 \geq 0$. The customer contract fulfillment (CCF) cost we discussed earlier is $[1-p]V$, which is decreasing in p. Thus, in the second-best contract, the intermediary will choose $[1-\hat{p}]$, the credit risk its customers are exposed to, by trading off $FC(p)$ against the CCF cost, as shown in *Figure 3* below.

Figure 3: The Exposure of the Customer to the Intermediary's Credit Risk in the Second-Best



4 Examples of Customer Contracts and Institutional Design

Many observed customer contracts and institutional designs fit within our framework. We briefly discuss some examples of these in this section.

4.1 Customer Contracts

4.1.1 Bank Deposits

A demand deposit in a bank represents a contract between the bank and a customer (depositor). In practice, the funds provided by depositors are invested in risky securities (e.g. loans), and uninsured depositors are exposed to the credit risk of the bank, consistent with the second-best contract. However, the depositor would prefer not to have to worry about the credit risk of the bank if the bank could find a cost-effective way to achieve this.

There are a number of potential institutional formats that achieve this. One is for the bank to be a “narrow bank”, i.e., a bank that invests only in assets like U.S. Treasury bonds that are free of default risk. For example, a money market mutual fund that invests in short-maturity U.S. government bonds would functionally be an example of a narrow bank. In this case, depositors are not exposed to any bank-specific credit risk, but such a bank would be reduced to its basic functions. Because of its limited scope, a narrow bank may be an extreme case, and the economic functions its scope would exclude may have to be performed less efficiently elsewhere in the economy, representing a potential economic loss.

A second way to do this is to have the bank raise financing from investors that is junior to the claims of depositors. This financing can come from investors who purchase the bank’s preferred stock and common stock. This allows the bank to create a cushion that absorbs some of the bank’s idiosyncratic risk and keeps it from affecting depositors. A similar role is played by third-party guarantors, whether they are private-sector guarantors or government deposit insurers. But as we pointed out in the previous section, it may be prohibitively costly to eliminate all of the customer’s credit risk exposure this way.

A third way to achieve the informational insensitivity that customers desire is to have deposit insurance. This insulates the depositors against the credit risk of the intermediary, even without narrow banking. However, deposit insurance has some well-known shortcomings. When deposit insurance is incomplete (so that only a portion of deposits are insured), some depositors may be exposed to the bank's credit risk if the bank makes risky investments. Indeed, no country has unlimited deposit insurance, so some deposits are uninsured.[32] Moreover, deposit insurance creates moral hazard (see Merton (1977)). Thus, even in the presence of deposit insurance, the second-best contract may expose the customer to some credit risk.

4.1.2 Mutual Funds

For mutual funds, customers are investors in the fund—each customer is purchasing a service (in this case, the portfolio management service and the promise of some risky return), while also providing financing through the money invested in the fund. In this case, the customer's final contract payoff need *not* be a riskless claim. Indeed, the customer understands that the contract purchased from the mutual fund may have a risky payoff, for example, linked to the S&P 500. It is only the credit risk of the intermediary that the customer wishes to be insulated from. That is, the customer does not want any uncertainty about receiving the portfolio management service or promised payoff because the fund is insolvent, possibly due to unobserved risky investments with fund money or "tunneling" that leads to a misuse of fund capital.

A mutual fund is a good example of a contract that imposes risk on the investor (customer) that is related to the contract itself.[33] As long as the mutual fund does not have any risk that the fund investors are exposed to beyond the risk of the underlying portfolio on which the fund

[32] It may also be that some deposit insurance is optimally kept incomplete, to the extent that some depositors could be considered as investors rather than customers. For example, they may be agents who have liquidity on which they seek a yield, but want to have the flexibility to withdraw it at a moment's notice in case some future (unanticipated) need for that liquidity arises. However, they have no need for any other service that the bank offers. Jumbo CDs and brokered deposits are examples of this. Unlimited deposit insurance may be made available only if banks in the aggregate are in trouble, as during the 2008-09 financial crisis.

[33] E.g. the custodian holds securities and provides insurance on theft.

investors have equity claims *and* there is no informational asymmetry between the fund manager and investors about the assets the fund is investing in, investors are willing to accept fund portfolio risk in their role as customers. Indeed, customers often invest large parts of their life-savings in mutual funds, with an expectation that they will be exposed to some degree of market risk, depending on what the fund invests in.

The desire to confine their risk exposure to only the (systematic) risk inherent in the fund's contractually-stipulated investment portfolio is one reason why investors put their money in funds managed by reputable intermediaries like Vanguard, Fidelity, and the like. That is, if one invests in the S&P 500 through one of these funds, the risk (R) is $R_{S\&P\ 500}$. Indeed, other than differences in expenses, the risk in the fund is the same regardless of whether it is offered by Vanguard or T Rowe Price or American Century. If these funds were rated, they would all be AAA, even though their future value is (systematically) risky. However, if one chooses to invest in the S&P 500 through an individual/company of lesser reputation, say agent *XYZ*, then the investor's risk is $R_{S\&P\ 500} + R_{XYZ}$, where R_{XYZ} is the credit risk of *XYZ*. Thus, in the case of many mutual funds, the second-best contract closely approximates the first-best as customers are exposed to little, if any, credit risk of the intermediary offering the fund.

4.1.3 Insurance Contracts

An individual who purchases a whole life insurance policy is a customer who is buying a bundle of two products—an insurance payoff in the event of death and an investment. The policyholder is willing to accept randomness in the return on the investment portion of the product, but not the risk that the insurance company may fail due to other exposures and hence be unable to pay in the event of the death of the insured. If the insured were to be exposed to such risk, it would represent inefficient risk bearing because it would not be efficient for the insured to buy a large number of smaller life insurance policies to diversify across insurance companies. The key, according to our framework, is in the degree of transparency for the customer. The efficient contract would be one that involves the insurance company designating a special account—analogous to the kind of bankruptcy-remote trusts or special investment vehicles used in securitization—in which the policyholders' funds are kept and then invested in a pre-designated basket of securities. That way the policyholders would not need to have any information about the credit risk in the rest of the insurance company. Absent this, an insurance fund that backs up insurance companies and protects policyholders would be needed. In all fifty U.S. states, state insurance funds provide this service.[34] With both arrangements, the policyholders would not need to have any information about the credit risk in the rest of the insurance company. Thus, the second-best contract achieves close to the efficiency of the first-best.

4.1.4 Repurchase Agreements (Repos)

Repos have been the mainstay of short-term financing in the shadow banking sector for over a decade, and this sector is now globally bigger than commercial (deposit-based) banking. A repo contract is an excellent illustration of our framework. The financial intermediary is an institution that has collateral in the form of bankruptcy-remote securities like U.S. Treasuries or high-grade mortgage-backed securities, but has need for liquidity over a short time period. The customer is another institution that has excess liquidity on which it wishes to earn additional yield income. The customer provides financing to the intermediary in exchange for taking ownership of the collateral for the duration of the loan. Since the loan amount is less than or equal to the value of the securities used as collateral, the customer is *not* exposed to the credit risk of the

[34] This also applies to property and casualty insurance. Property and casualty guaranty funds are part of a non-profit, state-based system that was created by statute, which pays outstanding claims of insolvent insurance companies. As of 2015, there were about 550 insolvencies since the inception of the guaranty funds.

intermediary. As a result, the second-best arrangement approximates the efficiency of the first-best.[35]

4.1.5 CoCo Bonds

As a way to reduce taxpayer losses from bank failures, a number of papers have proposed Contingent Convertible (CoCo) bonds as something banks should have in their capital structures (see Flannery (2013)) for a review of this literature. Under the Dodd-Frank Act, the FDIC has the power to recapitalize financially-distressed firms by “bailing-in” bondholders. However, such recapitalization when the firm is financially stressed can be challenging. CoCos, which are essentially convertible bonds that can be converted to equity via a regulatory trigger, provide an alternative way to recapitalize a distressed firm.[36]

While this is the main objective of a CoCo, our analysis provides another perspective on the role of CoCos. Imagine a bank with (pure) investors and customers. Most of the bank’s financing is in the form of debt, but some of the debt is provided by customers and some by investors. Then the investors can buy CoCos and the customers can buy more senior (non-convertible) debt. The CoCos would necessarily be exposed to the credit risk of the bank. However, a sufficiently large amount of CoCos can insulate the more senior debt from the bank’s credit risk, making it a good kind of debt for customers (even in non-depository institutions) to

[35] Indeed, whenever there are doubts about the extent to which the repo contract is insulated from the credit risk of the borrower, counterparties often refuse to enter into the contract or substantially increase the “haircut” on the repo. For example, as it approached insolvency, Bear Stearns was unable to find counterparties for even repos involving Treasuries as collateral.

[36] There is a large literature on how a regulatory trigger can itself serve as a self-fulfilling prophecy (in the sense of Merton (1948)) and add to the firm’s financial distress, especially when the conversion trigger is based on the market value of the bank’s equity. See, for example, Sundaresan and Wang (2015). A related issue is the knife-edge case in which a firm is close to the conversion threshold, leading to changes in behavior by market participants to try and cause conversion. This can be addressed in a straightforward fashion by more sophisticated designs. Potential solutions to these issues include making triggers mechanical and contractual, such as conversion when the book value of equity falls below 2.5% of assets, and the use of serial bonds or bonds that have the ability to reconvert to debt. Thus, the variety of triggers and designs possible with CoCos is extensive. In general, the value of CoCos is driven by conversion, and conversion is driven by credit quality, so the Merton (1974) model serves as a useful starting point for valuing these bonds.

hold. This may function in a similar way, from the view of the customer, to no-fault default or no-fault bankruptcy. This can move the intermediary's second-best contract with the customer close to the first-best in terms of efficiency.[37]

4.2 Customer Contracts and Institutional Design: Futures Exchanges

A futures contract essentially guarantees the ability to sell or buy some commodity or security in the future at a price that is predetermined. If this contract were negotiated as a forward contract with a financial intermediary, the holder of the contract would be the intermediary's customer.[38] Clearly, if the intermediary becomes insolvent prior to the delivery or execution date on the contract, the customer would be unable to avail of the insurance against the price risk that the customer sought under the contract.

A futures contract is traded on an exchange with liquidity and collateral provided daily, rather than being merely a bilateral arrangement between the bank and the customer that may not be collateralized. The exchange stands behind the execution of the contract. Consequently, the customer is protected against counterparty risk. Thus, the use of futures contracts over forward contracts may be rationalized as a means of insulating customers against the credit risk of an intermediary.

[37] See Merton (1992c) on no-fault default.

[38] With swaps and forwards, the two parties can switch back and forth between being a creditor or debtor to one another based on movements in the underlying assets. So approximately half the time, the customer will be owed money by the intermediary and thus be credit-sensitive. As noted previously, if a customer owes money to the intermediary, then he does not fit into the definition of credit-sensitive customers that we focus on here. In this sense, options traded on an exchange may be a better example. However, even for swaps/forwards, as long as there is a chance that the customer will become the creditor in the contract (if the market value of the underlying goes the other way), the effect that we emphasize will still be present.

5 Regulatory Practices

In this section, we apply our framework to two important regulatory practices: the recently-enacted Dodd-Frank Act, and the practice of protecting large banks against failure.

5.1 The Dodd-Frank Act

One aspect of the Dodd-Frank Act that can be understood within the context of our framework is that under Title VII of the Act, all non-exempt swaps to which a clearing exception does not apply (i.e. "standardized" swaps) must be cleared and exchange traded. Mandatory clearing and exchange trading of swaps is already underway. Our framework provides an economic rationale for this. By making swaps exchange-traded, counterparty credit risk is greatly reduced. Thus, the customers who hold these swap contracts need not worry about the credit risk of the intermediary they are working with, provided that the exchange is bankruptcy remote.

5.2 Too-Big-To-Fail

Firms and banks fail quite often, and these failures have negative consequences for at least some of their stakeholders. Typically, these negative externalities are not considered excessive and the usual resolution provided by the bankruptcy code is allowed to operate, leading to either a reorganization and restructuring of the firm or liquidation. However, in some cases, the negative externalities from the firm's failure and subsequent bankruptcy resolution are estimated to be so large that the government steps in and protects the firm's investors (creditors and sometimes even shareholders) from losses that would have been suffered if the usual bankruptcy resolution regime had been used. In this case, the firm is deemed to be "too big to fail" (TBTF). The concept of

TBTF is most commonly applied to banks and other financial firms—in such instances, the firm may be deemed a "systemically important financial intermediary (SIFI), and thus subject to additional rules and more oversight. One perhaps unintended consequence may be that such institutions may gain an advantage over others because the SIFI designation may signal a government *de facto* (implied) credit guarantee that non-SIFI institutions do not have.[39]

It is worth clarifying what we mean by TBTF. As Kaufman (2013) and Seelig (2004) point out, there are many types of TBTF interventions. The one we have in mind is the one in which the insolvent bank's shareholders are both protected against the loss that they would suffer in bankruptcy, and they remain in control of the institution, with the cost of this protection being absorbed by a third party. Since the shareholders are not wiped out, all customers and other creditors are fully protected against any loss. Most importantly for our purposes, the contracts the bank's customers have with it remain unruptured. Such a resolution is called "open bank assistance".[40]

There has recently been a surge of interest in TBTF, especially since the 2008-2009 financial crisis led to highly visible too-big-to-fail interventions, and these interventions are funded by the taxpayers. Because even in-the-money counterparties of failing firms are often protected, the taxpayer funding of such bailouts can be quite large. Moreover, institutions protected by TBTF may also exhibit moral hazard, investing in excessively risky assets (see Strahan (2013) and Afonso, Santos, and Traina (2014)). For these reasons, there is controversy about the desirability of TBTF.

[39] One could argue that a case such as General Motors (GM) might be close to TBTF as well, but for other economic reasons than the ones we emphasize here.

[40] For the purposes of our discussion, we would even include an intervention in which management does not remain in control, but all customers and investors are protected fully as the failing bank is acquired by a healthy bank, with assistance from the government. An example of this is the government-assisted acquisition of Bear Stearns by JP Morgan Chase during the 2008 financial crisis.

TBTF typically covers large and complex financial institutions. Small banks are routinely allowed to fail when they become critically undercapitalized.[41] This raises an interesting question that we address in this section: why are only large and complex financial institutions protected by too-big-to-fail, and small banks allowed to fail? The standard answer to this question is that the failures of large banks can result in contagion and cause the toppling of other banks. The specter of a systemic collapse is one reason why governments are interested in preventing large-bank failures. What we will argue is that TBTF protection for large institutions may make economic sense *even without* such systemic risk concerns. Moreover, our framework explains why too-big-to-fail is limited to large and complex financial institutions. Note first that too-big-to-fail is costly to taxpayers and is controversial, so the estimated benefit of the intervention has to be large enough to overcome the cost. The question thus boils down to why the benefit-to-cost ratio is viewed as being higher for large and complex institutions than for smaller ones.[42] Of course, leaving aside the issue of government bailouts, the reason why customers in both small and large banks are exposed to the credit risk of their banks is that second-best contracts are in place.

At a basic level, one explanation for the special treatment of large banks is that there are far more customers in large banks, so any disruption of the bank's provision of services to these customers—something that is unavoidable if the bank fails and especially if it is liquidated—means that more customers are exposed to its credit risk ex post, which represents a shock that was unanticipated ex ante. However, beyond this, we argue that there is greater *intertwining* of customers and investors in large banks, and hence it is more difficult to separate out customers and investors in large banks than in small banks. Given second-best contracts, this produces greater

[41] Of course, if the bank has insured deposits, these claims would be covered, but management would lose control and the bank's assets would be sold to another bank by the deposit insurance.
[42] Note that we are not referring to the absolute cost of the intervention to taxpayers, which is obviously larger for bigger banks (as is the benefit). We are referring to the cost-benefit ratio.

efficiency losses from failure for large banks than for small banks.

The essence of this argument is that the greater intertwining between customers and investors makes it very difficult to cleanly separate customers from investors, which then makes it challenging to protect customers from the credit risk of the intermediary while exposing investors to that risk. That is, this intertwining is likely to result in greater credit risk exposure for customers than would otherwise be the case with a second-best contract. As a result, a failure of a financial institution will be dysfunctional for the customers of that institution—when the institution goes bankrupt, the claims of investors and customers are settled in court, and it is uncertain what customers will receive. This amounts to a lottery for the customer when this happens, which we argued is inefficient.

There are many specific ways in which this effect could operate. Here, we provide a few examples to illustrate the main idea. A straightforward example is a bank that provides transaction services to depositors, and uses the liquidity provided by deposits to fund the services provided to customers. Some of the depositors are also investors in the bank. Now imagine that the bank is financially distressed, and is allowed to fail, wiping out all the investors. Because the investors have lost their money, they then rush to withdraw their deposits—even if these deposits are insured and hence safe—in order to meet their own liquidity needs and invest the money elsewhere in order to meet their investment goals. The loss of deposits then deprives the bank of the ability to provide deposit services, thereby adversely affecting the depositors who have not withdrawn their money. This kind of spillover effect is more likely in a larger and more complex bank, as this type of bank is more likely to have multiple types of investors who are intertwined with customers.

As another example, imagine a bank that has multiple customers, and which has arranged highly-customized swap contracts for a subset of its customers. Some of these customers are also investors in the intermediary, but given the size of the intermediary, it would be difficult and expensive for an outside party—even the bank's regulator—to determine which of the bank's customers are also investors and how much each has invested in the bank. If this bank fails and is not bailed out, the investors will experience a possibly significant decline in their wealth due to their losses on their investments in the bank. This wealth decline could lead some of these investors to discover that the swap contracts they had initially entered into are no longer useful. This may cause them to walk away from these contracts. If the bank has failed, it may not be in a position to step in as a counterparty to replace the lost customer. Thus, the bank's swap contracts may be adversely affected, with associated welfare losses.

A somewhat different example is an individual who has purchases a whole-life insurance policy from an insurance company that is a subsidiary of a bank holding company. This individual is also an investor in the bank holding company, holding some of its equity and bonds. If the bank holding company fails and is not rescued by the government under TBTF, its shareholders and bondholders are wiped out. This results in a significant decline in the wealth of the individual who has purchased the life insurance policy, and the decline may be so great that the individual cannot afford the premium on the policy anymore. He may therefore abandon the policy, thereby destroying the consumer surplus that was initially anticipated to be created by the policy. Thus, in highly-intertwined intermediaries, there are many ways in which customers can be adversely affected and social surplus destroyed as a result when investors are wiped out. The bigger and more complex the intermediary, the more difficult it is to protect the sheep while letting the wolves perish.

Such spillover effects from investors to customers due to the complex and difficult-to-detect intertwining of customers and investors in large financial intermediaries often compel regulators to bail out TBTF institutions, because their lack of knowledge of the extent of the intertwining makes them hesitant to see what happens if they do not bail out these institutions. Indeed, this was a rationale for the Federal Reserve's intervention to rescue Bear Stearns, given its rather significant role in the repo and swap markets, as indicated earlier. This demonstrates that, if one is unable to partition out customers and investors, then it may be necessary to guarantee investors in order to protect customers. Such concerns are substantially smaller in less complex and systematically less important (smaller) banks. Having said this, failures of even small banks are inefficient due to the impact of the failures on these banks' customers. But it may be that the costs of TBTF discussed earlier may exceed the costs of letting such banks fail.

Finally, while we have focused here on how our framework provides a way to understand bailouts, it can also help to analyze and better understand other related regulatory policies. One such policy that is also designed to relieve banks in distress is to have "bail-in deposits"—deposits that transform into equity when the bank is in trouble (see Zenios (2014) for a discussion of bail-in policies in the Cyprus crisis). Our analysis suggests that the optimality of having bail-in deposits depends on whether the depositors are customers or investors ex ante, and whether they can be separated. If certain depositors expect ex ante that they may be transformed into equity investors, then they can be considered to be investors. For these investors, a bail-in may make sense as a way to shore up banks. However, if depositors are customers as we have defined them, then they will be forced to be exposed to the bank's credit risk by becoming equity, which we have argued is inefficient (at least in a first-best sense). Therefore, a bail-in policy may make sense if customers and investors can be separated; otherwise, customers will be forced to become investors, thus making the policy inefficient.

6 Customers and Financial Crises

Our framework may provide some new insights into how financial crises are propagated, and how they should be dealt with. A *crisis* can be thought of as an event that is outside of agents' "model of the world"—when the event happens, agents have difficulty reconciling it with their view of the world, and therefore do not have a predetermined direction to go once it happens. For example, the most recent financial crisis was precipitated by a large decline in housing prices, which was an event that many people did not expect and did not have a contingency plan for. A defining element of a crisis is closely linked to the defining element of customers that we have described. When customers expect to receive a credit-risk-insensitive vector of services from contracts with intermediaries because they believe these are first-best contracts and unexpectedly learn that these contracts have been jeopardized by the impending insolvencies of these intermediaries, it can generate the same forces that give rise to a financial crisis—an unanticipated event which the customer did not consider a possibility and thus does not know how to respond to. Consequently, customers may withdraw their funds and cause a crisis. This correspondence between crises and customers also underscores how difficult it is to deal with such financial episodes, since customers, who provide a substantial portion of funding (but not risk-taking), attach different values to the services they receive under their contracts. This heterogeneity can lead to customer reactions to news about elevated intermediary credit risk that are quite unpredictable.[43] More directly, one of the reasons (even uninsured) intermediaries are so often bailed out by the government during crises is because investors and customers are intertwined in complex ways in many cases, and the unpredictability of customer reactions represents another

[43] Since the market for customer contracts is incomplete, the intertemporal marginal rates of substitution that are used to value future random payoffs are not the same across all customers.

reason why regulators are unwilling to "roll the dice" and let institutions fail.[44] Indeed, when such institutions go through bankruptcy, customers are inexorably caught up in the bankruptcy process, which makes it uncertain what claim they will eventually get through the courts.

A key to exploring this potential connection between customers and crises is to understand how a financial crisis may be propagated by the beliefs revision process of customers who experience breaches of contract. While in a typical setting, a new event would prompt economic agents to revise their beliefs in a Bayesian manner, for a customer who believes the contract is first-best, a breach of contract due to exposure to the credit risk of the intermediary is unexpected and is outside the "model of the world" —this precludes the usual Bayesian revision of beliefs. This can cause the customer to question the very model of the world he started with, and sometimes result in a switch to a different model. For example, a customer may have started with a model of the world in which its contract with the intermediary is inoculated against the credit risk of the intermediary. But when the customer finds out that the continued provision of intermediary services is predicated on the solvency of the institution and thus only second-best, it would switch to a different model. When this happens, the revision of beliefs is no longer a "within-model" Bayesian process, and this can introduce discontinuous, non-Bayesian shifts in beliefs. The result can be massive fund withdrawals by investors and a liquidity crisis for institutions that is triggered by concerns about the impact of their credit risk on the fulfillment of contracts with customers. In the Appendix, we provide one way to develop this argument more formally using recent work on non-Bayesian beliefs revision.

Our approach bears similarity to the "neglected risks" approach to financial crises in

[44] For example, money market funds, which fully disclose that they are not insured or guaranteed by anyone. However, in September 2008, the U.S. Treasury instituted an insurance program to prevent money market funds from breaking the buck.

Gennaioli, Shleifer, and Vishny (2012, 2015). They argue that investors often ignore (highly unlikely) tail risks, perhaps because they use intuitive short-cuts based on memory that tend to overemphasize the more salient risks (e.g. as in Bordalo, Gennaioli, and Shleifer (2012), and Gennaioli and Shleifer (2010)), leading to excess demand for some assets. However, when the risks in these assets become more evident, investors flee them, and this could precipitate a crisis. A key distinction is that our argument is focused on the aversion of customers to the credit risk of the intermediary as the risk most relevant in this calculus.

7 Conclusion

In this paper, we have developed the notion of "customers" and "investors" as a framing of the roles played by different groups of agents in funding financial intermediaries. Customers provide a significant amount of the funding, but want no risk-bearing. In contrast, investors provide both funding and risk-bearing. The customers' dislike for the credit risk of the intermediary makes them different from investors, and this distinction leads to a rich set of implications.

The most important implication that we focus on is the economic rationale for designing efficient (first-best) contracts that insulate customers from the credit risk of the intermediary and impose all of this idiosyncratic risk on the investors. We show that because customers cannot replicate the services they receive from intermediaries due to a form of market incompleteness, exposing customers to the idiosyncratic credit risk of the intermediary results in an inefficient loss of economic surplus. Intermediaries thus have an incentive to design contracts that protect the customer from the intermediary's credit risk. However, in a second-best setting in which providing such risk insulation is costly, a tradeoff must be made between the cost of insulating the customer from the credit risk of the intermediary and the cost of leaving the customer partially exposed. This perspective helps to explain the design of a variety of contracts in the real world—including not only deposit contracts in banking, but also the other contracts such as mutual funds, insurance contracts, repos, and CoCos. It provides a fresh perspective on why we have exchanges, and generates an economic rationale for the swaps clearinghouse requirement of the Dodd-Frank Act. It also helps to explain why too-big-to-fail protection is offered to large and complex banks, but not to small banks. Finally, this perspective also helps shed light on how financial crises can be propagated by the customer-investor nexus.

We view this framework as a useful starting point for identifying and understanding how the key roles of customers and investors impact financial intermediaries. Future work could derive additional equilibrium predictions in a more formal setting based on this framework. In addition, our framework also has implications for how certain types of contracts should optimally be structured, such as debt contracts between intermediaries and customers. An interesting extension would be to look at how our framework bears on the work related to opacity and transparency in contracts.

Appendix

A.1 A Formal Framework of Customers and Crises

To provide one way to more formally explore the link between customers and financial crises, we use the model developed by Ortoleva (2012), who provides the axiomatic foundations for non-Bayesian updating of beliefs in such an environment. While this phenomenon could be captured in other ways, we view this framework as a useful starting point in describing it. The basic idea is to introduce two types of uncertainties: (i) "within-model" uncertainty about future outcomes, and (ii) model uncertainty. Fixing the economic model the agent has for the determination of outcomes that the agent has in mind, there is a (subjective) probability distribution over those outcomes that captures the agent's prior beliefs, given that model. However, the agent may not be entirely sure of the correct model of the world. Model uncertainty means that the agent is unsure of the "correct" economic model that determines outcomes and thinks it could be one of $M > 1$ possible models. Thus, if $\tilde{x}$ is the random outcome whose future realization will be observed in the future, $g_i(x)$ is the probability density function of $\tilde{x}$, then model uncertainty means that the set of possible density functions (or economic models) is:

$$M = \left\{ g_i \mid i \in \{1, \ldots, M\} \right\} \tag{A-1}$$

The agent will have some prior beliefs, represented by a probability distribution, say π, over g_i. This distribution is called a "prior over priors". The main idea in Ortoleva (2012) is that the agent establishes some "threshold belief", say a probability $\varepsilon > 0$, and starts by picking a model of the world, say some g_i. The g_i chosen is the one that the prior over priors assigns the highest likelihood to, i.e., he picks the most likely model of the world in a maximum likelihood fashion. He then observes the realized value of $\tilde{x}$. If the prior probability of $\tilde{x}$ occurring, given g_i, is above ε, he uses Bayes rule to arrive at a posterior belief, without changing the model of the world he started with. But if the realized value of x had a prior probability of occurrence that was below ε, the agent would think that he may have chosen the wrong prior. In this case, he will update his prior over priors, and then pick the g_i that has the highest likelihood in the posterior belief over priors, given this updating. A new model of the world (g_i) may be chosen, after which Bayes rule is used to update beliefs related to within-model uncertainty.

To illustrate how this applies in the context of financial intermediaries and customers, suppose that there are two possible models of the world: *Model I* is one in which intermediaries have taken all possible actions to ensure that the payoffs to customers are independent of the credit risk of the intermediary itself, so the contract is first-best. In this case, the payoff to the customers has a two-state payoff: it is a high payoff with probability $q \in (0,1)$ and a zero payoff with probability $1-q$. *Model II* is one in which the customer's payoff depends on the credit risk of the intermediary with which the customer is contracting, i.e., the contract is only second-best. There are three types of intermediaries: good (*G*), medium (*M*), and bad (*B*). The prior beliefs over intermediary types are represented by $p_i, i \in \{G, M, B\}$, with $p_i \in (0,1)\ \forall\ i$. In this model, the customers get a high payoff with probability q_i when contracting with an intermediary of type *i,* with $q > q_G > q_M > q_B > 0$. Assume $q > p_M$, so the customers' share of the economic surplus is positive.

Suppose customers start with a prior over priors described by probability $\pi_1 \in (0,1)$ that *Model I* is the true economic model of the world and probability $1-\pi_1 \equiv \pi_2 \in (0,1)$ that *Model II* is the true model of the world. Let $\pi_1 > 0.5$, so customers will choose *Model I* when the economy opens. In this model, the probability of a zero payoff is $1-q$, and assume that this is a very small number. Now suppose customers observe a zero payoff. If $1-q < \varepsilon$, then they will question whether they started with the right prior beliefs about the model of the world. They will update their beliefs about the true model, using Bayes Rule:

$$\hat{\pi}_1 = \Pr(\text{true model is Model I} \mid \text{zero payoff}) = \frac{(1-q)\pi_1}{(1-q)\pi_1 + \sum_{i \in \{G,M,B\}} p_i[1-q_i]\pi_2} \tag{A-2}$$

The customers will now compare the posterior belief $\hat{\pi}_1$ to the posterior $1-\hat{\pi}_1 = \hat{\pi}_2$. Suppose $\hat{\pi}_1 < 0.5$. Then customers will switch their view of the right model of the world and adopt *Model II*. Given *Model II*, customers will arrive at posterior beliefs about the type of the intermediary they are dealing with.[45] Thus,

$$\Pr(\text{intermediary is type } i \mid \text{zero payoff}) = \hat{p}_i = \frac{(1-q_i)p_i}{\sum_{i \in \{G,M,B\}} p_i[1-q_i]} \tag{A-3}$$

for $i \in \{G, M, B\}$. If

[45] These beliefs were irrelevant if *Model I* was viewed as the correct model of the world.

$$\sum_{i \in \{G, M, B\}} \hat{p}_i q_i < p^* \tag{A-4}$$

then customers will view the total economic surplus from contracting with intermediaries as negative, and all financing from customers will vanish. Thus, a financial crisis can come about as customers' perceptions of risk rise in a discontinuous manner from a perceived probability of the good-payoff state dropping from q to $\sum_{i \in \{G, M, B\}} \hat{p}_i q_i$.

In the context of our previous analysis, the payoff uncertainty reflects the fact that the final payoff from a contract may be uncertain (e.g. a mutual fund), while the model uncertainty relates to the possibility that the customers' belief that contract payoffs are insensitive to the credit risk of the intermediary may be incorrect. Our analysis here shows that the model uncertainty can create a situation in which a financial crisis is triggered and propagated. An important ingredient in our explanation is the somewhat unique role of customers in financial intermediaries as significant providers of financing to these intermediaries. This means that when these customers are exposed to the credit risk of the intermediary, they can exhibit a discontinuous change in beliefs that causes them to withdraw from the market. The resulting loss in liquidity can trigger a crisis, as we saw with the drying up of liquidity in the repo market during the 2008-2009 crisis.

References

1) Afonso, Gara, Joao AC Santos, and James Traina. "Do 'Too-Big-To-Fail' Banks Take on More Risk?" *Economic Policy Review*, Forthcoming (2014).

2) Akerlof, George A. "The Market for “Lemons”: Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics* (1970): 488-500.

3) Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Salience Theory of Choice Under Risk." *The Quarterly Journal of Economics* 127, no. 3 (2012): 1243-1285.

4) Calomiris, Charles W., and Charles M. Kahn. "The role of demandable debt in structuring optimal banking arrangements." *The American Economic Review* (1991): 497-513.

5) Campbell, Larry, and John P. Wilson, “Financial Functional Analysis: A Conceptual Framework for Understanding the Changing Financial System”. Working paper, HSBC London, 2014.

6) Dang, Tri Vi, Gary Gorton, Bengt Holmström, and Guillermo Ordonez. *Banks as secret keepers*. No. w20255. National Bureau of Economic Research, 2014.

7) Diamond, Douglas W. "Financial intermediation and delegated monitoring." *The Review of Economic Studies* 51, no. 3 (1984): 393-414.

8) Diamond, Douglas W., and Philip H. Dybvig. "Bank Runs, Deposit Insurance, and Liquidity." *The Journal of Political Economy* 91, no. 3 (1983): 401-419.

9) Flannery, Mark J. "Contingent Capital Instruments for Large Financial Institutions: A Review of the Literature." *Annual Review of Financial Economics* 6 (2014), 225-240.

10) Froot, Kenneth A., and Jeremy C. Stein. "Risk management, Capital Budgeting, and Capital Structure Policy for Financial Institutions: An Integrated Approach." *Journal of Financial Economics* 47, no. 1 (1998): 55-82.

11) Gennaioli, Nicola, and Andrei Shleifer. "What Comes to Mind." *The Quarterly Journal of Economics* 125, no. 4 (2010): 1399-1433.

12) Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. "Neglected risks, financial innovation, and financial fragility." *Journal of Financial Economics* 104, no. 3 (2012): 452-468.

13) Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. "Neglected Risks: The Psychology of Financial Crises." *American Economic Review* 105, no. 5 (2015): 310-14.

14) Gorton, Gary, and Guillermo Ordoñez. "Collateral Crises." *American Economic Review* 104,

no. 2 (2014): 343-78.

15) Gorton, Gary, and George Pennacchi. "Financial Intermediaries and Liquidity Creation." *The Journal of Finance* 45, no. 1 (1990): 49-71.

16) Hart, Oliver. "Corporate governance: some theory and implications." *The Economic Journal* 105, no. 430 (1995): 678-689.

17) Hayek, F. A. "The Use of Knowledge in Society." *The American Economic Review* 35, no. 4 (1945): 519-530.

18) Hirshlelfer, Jack. "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review* 61, no. 4 (1971): 561-574.

19) Holmstrom, Bengt. "Moral Hazard and Observability." *Bell Journal of Economics* 10, no. 1 (1979): 74-91.

20) Holmstrom, Bengt. "Understanding the Role of Debt in the Financial System." BIS Working Papers No. 479. Bank for International Settlements, 2015.

21) International Bank for Reconstruction and Development, *Global Survey on Consumer Protection and Financial Literacy: Oversight Frameworks and Practices in 114 Countries*, World Bank, 2014.

22) Kaufman, George G. "Too Big to Fail in Banking: What Does it Mean?." *Journal of Financial Stability* (2014).

23) Kyle, Albert S. "Continuous Auctions and Insider Trading." *Econometrica: Journal of the Econometric Society* (1985): 1315-1335.

24) Merton, Robert C. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *The Journal of Finance* 29, no. 2 (1974): 449-470.

25) Merton, Robert C. "An Analytic Derivation of the Cost of Deposit Insurance and Loan Guarantees: An Application of Modern Option Pricing Theory." *Journal of Banking & Finance* 1, no. 1 (1977): 3-11.

26) Merton, Robert C. "On the Application of the Continuous-time Theory of Finance to Financial Intermediation and Insurance." *Geneva Papers on Risk and Insurance* (1989): 225-261.

27) Merton, Robert C. "Financial Intermediation in the Continuous-Time Model", in *Continuous-Time Finance*, 428-471. Blackwell: 1992a.

28) Merton, Robert C. "On the Cost of Deposit Insurance When There are Surveillance Costs", in *Continuous-Time Finance*, 501-511. Blackwell: 1992b.

29) Merton, Robert C. “No-Fault Default: A Possible Remedy for Certain Dysfunctional Consequences of Corporate Leverage”, Presented at the *American Finance Association Meetings*, 1992c.

30) Merton, Robert C., “Operation and Regulation in Financial Intermediation: A Functional Perspective”, in *Operation and Regulation of Financial Markets*, ed. Peter Englund, (The Economic Council, Stockholm, 1993), 17-67.

31) Merton, Robert C. "A Functional Perspective of Financial Intermediation." *Financial Management* 24, no. 2 (1995): 23-41.

32) Merton, Robert C. "A Model of Contract Guarantees for Credit-Sensitive, Opaque Financial Intermediaries." *European Finance Review* 1, no. 1 (1997): 1-13.

33) Merton, Robert K. "The Self-fulfilling Prophecy." *The Antioch Review* 8, no. 2 (1948): 193-210.

34) Merton, Robert C., and Zvi Bodie. "A Conceptual Framework for Analyzing the Financial Environment." Chap. 1 in *The Global Financial System: A Functional Perspective*, by D. B. Crane, K. A. Froot, Scott P. Mason, André Perold, R. C. Merton, Z. Bodie, E. R. Sirri, and P. Tufano, 3–31. Boston: Harvard Business School Press, 1995.

35) Merton, Robert C., and Zvi Bodie. "Design of Financial Systems: Towards a Syntheses of Function and Structure." *Journal of Investment Management* 3, no. 1 (2005): 6.

36) Ortoleva, Pietro. "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News." *The American Economic Review* 102, no. 6 (2012): 2410-2436.

37) Ramakrishnan, Ram TS, and Anjan V. Thakor. "Information reliability and a theory of financial intermediation." *The Review of Economic Studies* 51, no. 3 (1984): 415-432.

38) Seelig, Steven. "Too Big to Fail: A Taxonomic Analysis." *Too Big To Fail: Policies and Practices* (Praeger Publishers) (2004): 219-230.

39) Strahan, Philip E. "Too big to fail: Causes, consequences, and policy responses." *Annual Review of Financial Economics* 5, no. 1 (2013): 43-61.

40) Sundaresan, Suresh, and Zhenyu Wang. "On the Design of Contingent Capital with a Market Trigger." *The Journal of Finance* 70, no. 2 (2015): 881-920.

41) Wakker, Peter, Richard Thaler, and Amos Tversky. "Probabilistic Insurance." *Journal of Risk and Uncertainty* 15, no. 1 (1997): 7-28.

42) Zenios, Stavros A., “Fairness and Reflexivity in the Cyprus Bail-In” (August 2015).

Forthcoming, *Empirica, Journal of European Economics*; The Wharton Financial Institutions Center Working Paper No. 14-04.