Customers and Investors: A Framework for Understanding Financial Institutions[*]

Robert C. Merton[†] and Richard T. Thakor[‡]

This Draft: November, 2016

Abstract

Financial institutions are financed by both investors and customers. Investors, such as shareholders and bondholders or private/public-sector guarantors of institutions, expect an appropriate risk-adjusted return in exchange for providing financing and risk bearing. Customers, in contrast, provide financing in exchange for specific services, and want the fulfillment of these services to be free of the intermediary's credit risk, even when these customers 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 develop a theory that provides an economic foundation for the aversion of customers to intermediary credit risk. We have the following main results. First, with positive net social surplus in the bank-customer relationship, the efficient (first best) contract completely insulates the customer from the intermediary's credit risk, thereby exposing the customer only to the risk inherent in the contract terms. Second, when intermediaries face financing frictions, the second-best contract may expose the customer to some intermediary credit risk, generating "customer contract fulfillment" costs. Third, we provide a formal analysis of insurance contracting as a specific illustration of our theory. Fourth, we propose that the efficiency loss associated with these costs in the second best rationalizes government guarantees like deposit insurance even when there is no threat of bank runs. We further discuss the implications of this customer-investor nexus for a host of issues related to how contracts between financial intermediaries and their customers are structured and how risks are shared between them, ex ante efficient institutional design, and regulatory practices.

Key words: Customers, investors, credit risk, financial intermediaries, real-world financial contracts, information-insensitivity, financial crises

JEL Classification Numbers: D81, D83, G01, G20, G21, G23, G28, H12, H81

[*] Original Draft: June, 2015. We thank an anonymous referee, Doug Diamond, Zhiguo He, Stavros Zenios, and participants at the Federal Reserve Bank of Cleveland and the Office of Financial Research Financial Stability Conference for helpful comments.

[†] MIT Sloan School of Management and NBER

[‡] University of Minnesota, Carlson School of Management

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 means. A concrete example is an investment bank that considers participating in a bridge loan to start a merchant banking business, and in doing so risks having institutional customers flee its over-the-counter derivatives business (e.g. long-dated swap contracts) because of concerns about the bank's ability to fulfill its contractual obligations on its derivative products were it to suffer losses on its bridge loans (see Merton (1997)).[2] In financial intermediation theories, 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 theoretically examine how this aspect of the relationship between a financial intermediary and its customers affects the design of contracts between them, and how it illuminates commonly-observed features in various real-world contracts, institutions, and regulatory practices. In doing so, we build on Merton's (1989, 1993, 1995, 1997) insights, but go beyond them in providing a formal analysis of efficient contractual arrangements as well as

[1] See also Merton (1992a). These papers are part of the "functional perspective" of financial intermediation; see also Merton (1990).

[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.

deviations from efficiency due to contracting frictions.

The starting point of our analysis is that financial institutions differ from non-financial firms in at least two noteworthy respects. First, the investors in a financial institution purchase claims that look similar to what its customers purchase. For example, a bank's depositors are customers who purchase liquidity and transaction services (see DeAngelo and Stulz (2015)) and have a debt claim on the bank's cash flows, whereas its subordinated debtholders are investors who also hold debt claims on the bank. By contrast, customers in a non-financial firm like IBM purchase products that are transparently different from the claims of its investors. Second, in financial institutions both investors and customers provide financing to the intermediary.[3] Investors, like shareholders and bondholders, provide financing and risk bearing since the values of their claims are linked to the intermediary's outcomes. Customers, in contrast, expect services in exchange for the financing they provide, but prefer not to bear intermediary-specific credit risk, i.e., they want the intermediary's service provision to not depend on the fortunes of the service provider.[4]

There are two types of customers in financial intermediaries that we distinguish between: "credit-sensitive" customers and "other" customers. "Credit-sensitive" customers provide financing *to* the intermediary in exchange for future intermediary services. The financing provided by these customers is a liability of the intermediary. The utility customers derive from the intermediary's services is diminished by an increase in the credit risk of the intermediary. "Other" customers are those who receive financing *from* the intermediary, such as bank borrowers. They

[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.

[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.

appear on the asset side of the intermediary's balance sheet, and are *not* credit-sensitive since they have an obligation to repay the intermediary in the future. Our focus 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 investors to bear it. Thus, a financial intermediary that imposes credit risk on its customers will not be able to compete effectively 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. This sheds light on some survey evidence. 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 rose from 0% to 1%. Wakker, Thaler, and Tversky (1997) argue that this is hard to reconcile with standard expected utility theory. We provide a rational explanation for such behavior.

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, the agent may play multiple roles, and may therefore have different expectations of the institution in different roles, e.g., a policyholder in an insurance company is a customer but may also hold the company's stock as an investor. 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.[5]

The questions we address in this paper are: what are the implications of this customer-investor

[5] For example, an individual will be a customer of a bank in which he/she has a retail deposit account and an investor with respect to the purchase of stocks of publicly-traded firms.

nexus for how the financial intermediaries structure efficient (first-best) contracts with their customers? That is, *why* do customers not wish to be exposed to intermediary credit risk? When financing frictions impede the adoption of efficient contracts, how does this perspective illuminate the microfoundations of observed (second-best) contracts between intermediaries and their customers? What implications does this have for certain institutional arrangements and regulations?

Our main results can be summarized as follows. First, we analyze the efficient (first-best) contract between the intermediary and the customer and show that as long as the contract creates positive net social surplus, it is structured so that the customer is completely insulated from the credit risk of the intermediary. Consequently, the customer is exposed only to the risk stipulated in the contract terms, and not the credit risk of the intermediary itself.[6] We show that exposing the customer to the intermediary's credit risk is akin to affixing to the contract a lottery that has negative social value, and that because of this all of the intermediary's credit risk is borne by its investors in the efficient contract. We further show that asking the customer to diversify exposure to the intermediary's credit risk by purchasing contracts from a large number of intermediaries is inefficient relative to the intermediary's investors bearing this risk. A key element of the argument is that the customer operates in an inherently incomplete market while purchasing a contract from a financial intermediary. However, our argument does not rely on any lack of sophistication on the part of customers or constrained access to information—the customers in our analysis are not simply "widows and orphans" or unsophisticated investors. A customer could be an institution

[6] As an example, 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.

such as the World Bank or a large pension fund.

Second, we analyze the second-best contract, which is constrained-efficient in the sense that the intermediary may face costly financing frictions that obstruct its ability to completely insulate the customer from the intermediary's credit risk. 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. The second-best contract may thus expose the customer to some of the credit risk of the intermediary, absent government intervention.[7] Indeed, some government intervention may be rationalized by the goal of reducing these costs. The loss of efficiency in the second-best is referred to as "customer contract fulfillment" (CCF) costs.

Third, in order to fix ideas in a specific context, we model an insurance company dealing with its customers (policyholders). We formally characterize the first-best and second-best contracts, prove that it is inefficient for customers to diversify across many insurance companies in order to diminish their exposure to credit risk, and explicitly derive the CCF cost. An interesting point emerging from this analysis is that in the second best, limited exposure of customers to *some* of the insurance company's credit risk can serve as a form of market discipline.

Fourth, we discuss how our analysis explains a variety of observed real-world contracts, institutions, and regulatory practices. The contracts that are rationalized by the framework developed in this paper are: insured bank deposits, mutual funds, insurance contracts, and repos in shadow banking.[8] An institution we analyze is a futures exchange. Our analysis offers insights into some regulatory practices in banking, specifically the Dodd-Frank Act, a major financial services regulation enacted in 2010 in response to the 2008-2009 global financial crisis. The element of

[7] This explains why customers are sometimes willing to deal with institutions that do not have a AAA credit rating.
[8] For mutual funds, the exception is if the mutual fund is providing liquidity services for cash.

this regulation that we provide economic foundation for is the requirement for swaps to be traded through clearing houses and exchanges.

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 to develop a theory that enables a characterization of the first-best contract. We introduce financing frictions for the intermediary in Section 4 to explain why such separation between the contract and the credit risk of the intermediary can be less than perfect in the second-best contract. We show how the optimal degree of exposure of the customer to the credit risk of the intermediary in the second-best contract is determined and how this generates CCF costs. Section 5 takes the framework to the specific application of an insurance company contracting with policyholders and solves for the first-best and second-best contracts. Section 6 turns to a discussion of how the analysis illuminates observed contracts, institutions, and regulations. Section 7 provides concluding remarks.

2 Related Literature and Marginal Contribution

In this section, we review the related literature and indicate our marginal contribution relative to this literature.

Our paper is most closely related to a series of papers by Merton (1989, 1993, 1997), so we begin by noting here the marginal contributions of our paper relative to those papers. First, Merton's analysis focuses on the efficient (first-best) contract. We formally analyze this contract and characterize its properties, but we go beyond it and model the financial frictions that may result in the contract not always being encountered in practice. In doing so, we explicitly characterize the second-best contract and describe the resulting loss in efficiency as a CCF cost. The characterization of the CCF cost is novel to this paper. Second, our analysis of an insurance contract with policyholders provides a specific and formal illustration of the key ideas in our framework. Finally, we explain how our analysis can shed light on numerous contracts, institutions, and regulatory practices. For example, it rationalizes federal deposit insurance even if there is no threat of bank runs.

Our paper is also related to the literature on the functional perspective in finance (for example, Merton (1990, 1993, 1995), and Merton and Bodie (1995, 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 examine 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.

Another related strand of the literature is that which offers reasons for why bank deposit contracts are designed to be information-insensitive. Gorton and Pennacchi (1990) first proposed that agents who lack the skills to efficiently acquire and process information would prefer to invest in instruments that are informationally insensitive so as not to be disadvantaged in trading with informed agents. They explain the demand for informationally-insensitive assets like riskless bank deposits on this basis.

Since then, others have rationalized debt contracts that are optimally informationally-insensitive for a variety of reasons, including optimal risk sharing. For example, Dang, Gorton, Holmstrom, and Ordonez (2014) rely on the Hirshleifer (1971) notion that information may sometimes not be released because its release can distort risk sharing.[9] In their model, there are two generations of depositors, who are effectively risk averse.[10] The early generation of depositors will wish to sell their claims to the late generation if hit by a liquidity shock, but they do not want the late depositors to produce information about the value of the bank's assets, since this makes their exit price information-contingent 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 discourage information production by the second generation of depositors. Thus, the bank in this setting should optimally be opaque—it should withhold information about the bank's risk from depositors, making deposits information-insensitive.

A somewhat different perspective on why bank deposits are (optimally) riskless, at least asymptotically, is provided by earlier work that provided the information-based microfoundations for financial intermediary existence. 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 projects), so that even if an individual loan that is monitored/screened by the bank remains (partially) opaque, the bank itself becomes riskless as it grows to its efficient size. The optimality of such an intermediary does not depend on depositor risk aversion, however.[11]

Not relying on risk aversion to explain the demand for safe debt is also consistent with the stylized fact that investors are willing to pay a "premium" for riskless debt by accepting a lower yield than implied by risk aversion (e.g. Krishnamurthy and Vissing-Jorgensen (2012)). A number of recent theories emphasize the special role of banks in liquidity creation and assign a liquidity

[9] See also Holmstrom (2015).

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

[11] 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.

premium to safe debt. DeAngelo and Stulz (2015) use this to explain why banks rely on risk management to minimize asset risk and maximize their deposit-holding capacity. Their analysis focuses on how this liquidity production induces banks to have such high leverage. Hanson, Shleifer, Stein, and Vishny (2015) present a model in which banks that hold risky and illiquid loans create safe debt claims by relying on deposit insurance and equity capital. This makes deposits informationally-insensitive and depositors indifferent to the bank's credit risk. Hart and Zingales (2014) theorize that banks represent a cost-effective way to "manufacture" safe debt.

These papers focus on the demand for safe debt. A supply-side perspective appears in the security design literature in which firms engage in tranching their total cash flows to produce a multitude of claims, some that are less information-insensitive than the total cash flows and others that are more information-sensitive than the total cash flows. For example, Boot and Thakor (1993) develop a theory of security design in which riskless debt and information-sensitive equity emerge as optimal contracts for an issuer that wants to maximize expected revenue from issuing securities. Their model also explains asset pooling, securitization, and tranching. DeMarzo and Duffie (1999) also develop a model in which an issuer has an incentive to raise capital by securitizing part of its assets. The issuer's private information at the time of security issuance causes illiquidity in the security. The security design problem trades off the cost of retaining cash flows that are excluded from the claims sold to investors against the cost of including these cash flows in the design of the securities sold to investors.[12] The paper characterizes conditions under which standard debt is an optimal security.[13] Similarly, DeMarzo (2005) shows that tranching allows the issuer to create a

[12] Their theory raises the issue of what originators of debt claims keep on their balance sheets and what they choose to let others hold. Erel, Nadauld, and Stulz (2014) document that banks that had greater securitization activity had higher holdings of highly-rated tranches.

[13] Fulghieri and Lukin (2001) examine optimal security design and show that the Myers and Majluf (1984) pecking order aversion of firms to equity need not hold when outside investors can produce information about the firm and the equilibrium degree of information asymmetry is endogenous. That is, they provide an information-based rationale for equity, rather than safe debt.

low-risk, liquid security. Thus, rather than focusing on customers' needs for contracts that protect them from the credit risks of the institutions they deal with, this literature focuses on how the creation of safe securities via tranching serves security issuers.

Our theory differs from this previous work in a number of ways. First, we draw a sharp distinction between customers and investors in financial institutions, and show that it is *only* the customers who should be protected from the fortunes of the intermediary in an efficient contract. Second, in our framework, it is not only bank deposits that should be optimally insulated from bank credit risk—and hence made insensitive to bank-specific information—but *all* efficient contracts between the financial intermediary and its *customers* that should be insulated. This includes a far bigger set of contracts and institutions than bank deposits. For example, insurance contracts, repos, and futures exchanges are also included.[14] Third, 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 state-contingent payoffs of the contract itself and *cannot* include the credit risk of the intermediary. Thus, we are not just talking about "safe" debt. Fourth, we address the important question of *why* all of the credit risk and the affiliated informational risk should be borne by the intermediary's investors in the first-best case, and not by its customers. This enables us to shed new light on issues like the need for deposit insurance even in the absence of the threat of contagious bank runs and the Dodd-Frank Act. Fifth, 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 being prohibitively 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

[14] Hanson, Shleifer, Stein, and Vishny (2015) also examine repos as an example of (non-depository) safe debt.

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. Finally, our analysis of the second best highlights the potential channels through which financing frictions can diminish efficiency in the customer-intermediary contract by imposing intermediary-specific credit risk on the customer, and the CCF costs generated as a result.

3 Financial Intermediaries and Customers

3.1 Analytic Setting: Efficient Customer Contracts (First-Best)

We now introduce a simple analytic example to define and discuss the key concepts concretely. 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 equivalent of the 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 the (possibly state-contingent) money F that is paid by the intermediary to the customer as part of the service provided by the intermediary. In a bank, F could be the amount of deposits (plus interest) withdrawn by the depositor.[15] In an insurance context, f_t would represent the vector of insurance premia paid to the insurance company and F 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 customer—it will capture the opportunity cost for the customer to use the financial intermediary rather than purchase the service through, say, another intermediary or even the financial market.[16] Satisfaction of the

[15] 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.

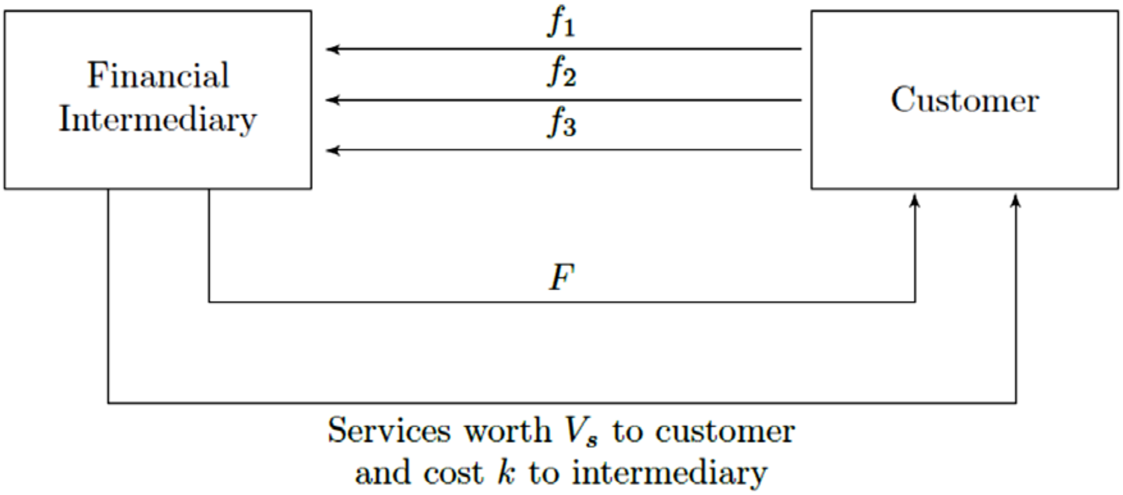
[16] Merton (1989) suggests one reason why an intermediary may be able to improve upon the market in providing service to the customer, specifically by providing *customized* derivatives securities that generate a payoff stream that replicates the customer's desired payoff stream emerging from an intertemporal portfolio optimization. Because the

customer'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.

Figure 1: Values and Costs of Financing and Services


Financial Intermediary
f_1
f_2
f_3
Customer
F
Services worth V_s to customer and cost k to intermediary

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

intermediary can aggregate derivatives contracts and then hedge risk in the market, the arrangement is more efficient than the individual customer trading directly in the market.

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

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 $pV + V_m^I - [\overline{V} + k]$.[17] Absent 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.

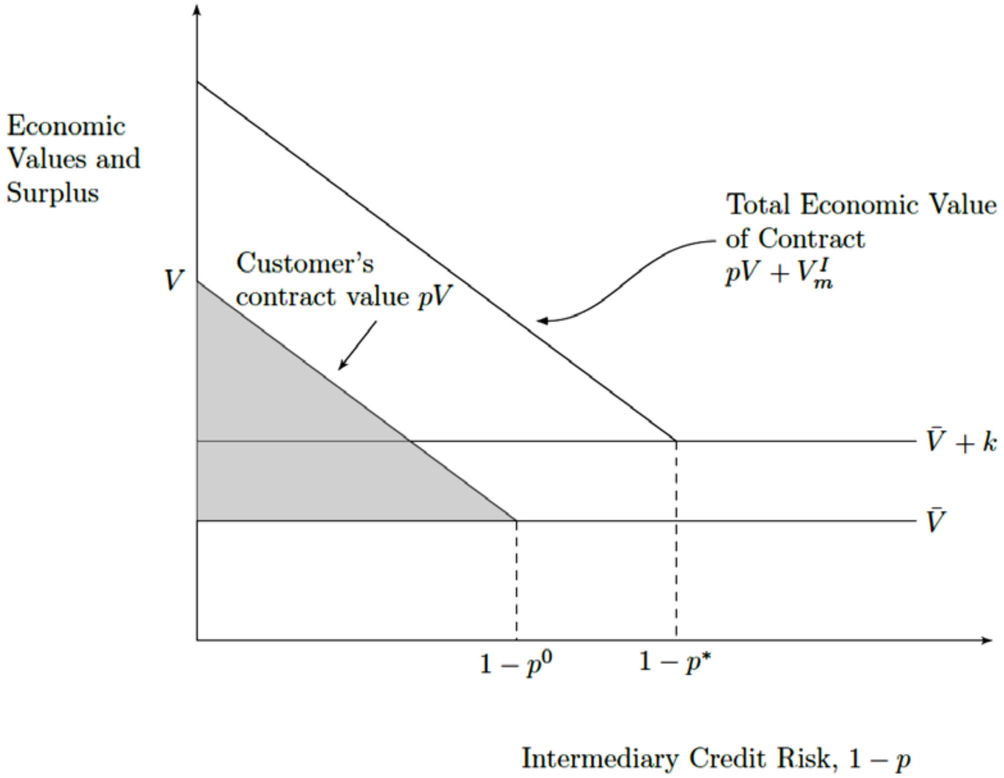
[17] V_m^I is not multiplied by p in these expressions because all financing is provided by the customer up front at $t = 0$. Thus, insolvency on the part of the intermediary at a later date will not reduce the value to the intermediary of obtaining financing from the customer.

We can now characterize how the total economic value of the contract surplus (to the customer and the intermediary) and the customer's share of this total contract value behave as functions of the intermediary's credit risk, $1 - p$. These relationships are depicted graphically in *Figure 2*.

Theorem 1: *The rate at which the total expected net economic surplus,* $ES_{total} = pV + V_m^I - [\overline{V} + k]$, *declines with respect to financial intermediary credit risk is the same as the rate at which the customer's net expected economic surplus,* $ES_c = pV - \overline{V}$, *declines with intermediary credit risk, and this rate is increasing in V. The intermediary credit risk,* $1 - p^0$, *at which* ES_c *becomes zero is less than the credit risk,* $1 - p^*$, *at which* ES_{total} *becomes zero. Moreover,* p^* *is decreasing in* $V_m^I - k$, *the spread between the monetary value of the intermediary's service and the cost of providing that service.*

This result implies that the larger the value of the service provided by the intermediary, the faster is the rate of decline of the economic surplus from the customer relationship due to an increase in intermediary credit risk, i.e., more valuable relationships are more sensitive to intermediary credit risk. The intuition for why the value of the customer's net expected economic surplus becomes zero at a lower level of intermediary credit risk than the level at which total expected net economic surplus becomes zero is that there is also a net producer surplus, $V_m^I - k$, for the intermediary (see *Figure 2* below). Since p^* is decreasing in this surplus, the larger this surplus, the bigger is the spread between $1 - p^*$ and $1 - p^0$.

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 $t = T$.[18] Second, this

[18] 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 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.

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; we take up this issue in Section 4.

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 the 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 related to 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 service-delivery contracts the customer has with the financial intermediary.

3.2 The Inefficiency of Exposing Customers to Intermediary Credit Risk

We now provide a microfoundation for the CCF cost discussed in the previous section. This analysis should be viewed as a specific example of how the CCF cost can be microfounded, and not the only way that economic surplus can be destroyed by exposing the customer to the intermediary's credit risk.

Consider a situation in which a financial institution raises financing from both customers and investors. The customers can be either risk averse or risk neutral. Since our focus is on the idiosyncratic credit risk of the institution, the assumption of investor risk neutrality is without loss of generality because they can diversify away the credit risk. Financing from customers occurs because customers essentially "pre-pay" for future services, as described in the set-up in the previous section. For example, the customers of an insurance company purchase insurance and pay premia for possibly many periods before they experience an accident or some other contingency they have insured themselves against, a feature that is an essential element of the way insurance works and how insurance companies finance themselves. This timing of the service provision at a *future* date exposes customers to the institution's risk of failure.

Assume now that feasibility of service provision to the customer at $t = T$ depends on information about the customer, represented by $\omega \in \Omega$ (where Ω is the feasible information set), and this information is privately acquired by the institution at a cost. For notational parsimony, let k represent the sum of the cost of this information acquisition as well as the cost of providing service to the customer. The intermediary's relationship with the customer generates proprietary information for the intermediary that its competitors lack, following the insights of the relationship banking literature.[19] Obtaining the incumbent intermediary's private information about the customer may either be infeasible for a competing intermediary or may cost considerably more, say $k_c \in (k,\infty]$. Suppose that for $\omega \in \Omega_s$ the institution can profitably provide the service, and for $\omega \in \Omega_f$ it cannot, where $\Omega_s \cup \Omega_f = \Omega$. If all that the institution knows is $\omega \in \Omega$, it is not profitable to provide the service at $t = T$.[20] This now leads to the following result:

[19] See Boot (2000) for a review of the literature.

[20] There are many examples of this. An insurance company must evaluate the health of its customers before deciding whether to sell a life insurance policy and at what price. Similarly, a bank will privately acquire information about its depositors through the cash inflows and outflows in their accounts and this may be useful information for determining

Theorem 2: *In the first-best case in which the institution faces no frictions in raising external financing, the contract between the institution and its customers completely protects customers from the credit risk of the institution related to its insolvency probability* $1 - p$.

To see the intuition, note that if customers are risk averse, then clearly the efficient contract completely insulates them from the risk of the institution and lets the risk neutral investors bear it. If customers are risk neutral, the risk sharing argument does not apply. But in this case, what is relevant is that the institution acquires private information about the customer that is pertinent to the determination of whether future services will be provided to the customer. Due to this, the failure of the institution is costly because it can deny service to the customer that would have been delivered had the institution been solvent, or the service may be provided at a higher cost.

3.3 Why is it Inefficient for Customers to Mitigate Intermediary Credit Risk?

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 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 its single-intermediary contract with a large number of smaller contracts with many intermediaries. However, one reason why we have financial intermediaries is that they

whether to grant a loan and at what terms, and if the loan is granted, further proprietary information is generated for the bank.

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 be inherently inefficient due to duplicated costs of information acquisition and service provision.

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 components 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 typically inseparable from the monetary component of the contract, as explained in Section 3.1.[21]

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 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, as suggested by the relationship banking literature, for example. This rules out a complete market in which state-contingent claims can be created with values that depend only on

[21] Moreover, 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.

states of the world and not on the “institutional affiliation” of each claim.[22]

One could argue that one way to circumvent this institutional affiliation 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 and Adverse Selection as Sources of Market Incompleteness

The third-party resolution mentioned above can generate *moral hazard* (e.g. Holmstrom (1979)) in the following way. 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 allocating their savings across multiple mutual funds managed by the company.[23] 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. When the intermediary’s provision of services to the customer and the continued financing of the intermediary by the customer are bundled into a single contract, there is a built-in incentive on the intermediary’s part to not underinvest in service provision. The third-party resolution encounters moral hazard by decoupling

[22] 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.

[23] For example, Fidelity and Vanguard give their customers the ability to split their investments across dozens of funds.

these two aspects of the contract.

The third-party resolution also suffers from potential adverse-selection problems. As mentioned earlier, the very nature of the intermediary-customer relationship is such that it generates proprietary customer-specific information. If a third party were to purchase the services the customer needs and then provide them to the customer, the quality of the service may be difficult to ascertain because of the superior information of the service provider about this quality and incentives to behave strategically, which could lead to a potential market breakdown (e.g. Akerlof (1970)).

4 Financing Frictions and the Second-Best (Constrained-Efficient) Contract

In our discussion of the first-best case, we assumed that the intermediary faced no financing frictions. In the absence of such frictions, all of the intermediary’s credit risk is efficiently borne by its investors and none by its customers. But we know from Myers and Majluf (1984) that adverse selection can make external finance costly relative to internal finance. In this section, we discuss the implications of this financing friction for the extent to which intermediaries choose to protect their customers from their own credit risks. That is, we analyze how financing frictions can cause the second best to deviate from the efficient (first-best) contract.

4.1 Adverse Selection, Investors, and Customers

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. For example, *p* may vary in an *a priori* unobservable way across intermediaries. 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 discovering the intermediary's credit risk becomes utility-relevant for the customer. 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 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. 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 each intermediary knows its own p privately. Others only know that p is distributed in the cross-section according to a probability distribution H. Then, if all financial intermediaries offer the same (pooling) contract to customers, they will value it at

$$\int Vp\, dH = \overline{p}V \tag{4}$$

And for $\overline{p} < p^*$, no contracting will occur because the total economic surplus will be negative. This creates an incentive for the intermediary to insulate its customers from its own credit risk.

This insulation, if it can be achieved, will succeed in allowing intermediaries to raise financing from customers. Merton (1995) discusses three ways in which the intermediary may attempt to achieve this insulation: by matching asset payouts with liability payouts, by purchasing third-party guarantees, and by keeping sufficient equity capital. However, when there is adverse selection, each of these approaches involves *transferring* the adverse selection problem to *investors*, i.e., the more the intermediary insulates the customer from adverse selection, the more of the adverse selection is shifted by it to the investors or the guarantor. 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, 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*?

An economic rationale for this lies in the fact that securities are *traded* in the capital market, whereas the claims of customers are *not*. 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.[24] This will impede the tradability of these contracts. 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.[25] This preserves the incentives for

[24] 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).

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

some investors to become informed, and also results in some of the adverse selection problem being dissipated.[26] 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—but not eliminate—the adverse-selection costs that intermediaries face in raising capital from investors.

The fact that investors' claims are traded also means that they can diversify away the idiosyncratic risk of the intermediary.[27] In an efficient market, each risk should be borne by those with the greatest risk-bearing capacity for that risk. Since customers do not trade their contracts, they cannot diversify away the intermediary's idiosyncratic credit risk. This tradability distinction tilts risk bearing towards investors.

4.2 The Effect of Financing Frictions on the Second-Best Contract

The discussion above indicates that even the second-best contract will seek to protect customers against the credit risk of the intermediary. Nonetheless, we cannot rule out the possibility that financing frictions and related costs faced by intermediaries may be so large that the second-best contract may involve the customer bearing *some* of the credit risk associated with

[26] 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.

[27] Pooling securities across intermediaries and then tranching the pool to create information-insensitive securities, as is done routinely with securitization, is another way to diversify away idiosyncratic risk (see Boot and Thakor (1993) and DeMarzo and Duffie (1999)). We would however, view the purchasers of these asset-backed securities (ABS) as investors rather than customers who purchase a bundled claim to a financial return and intermediary services.

the intermediary, i.e., the second-best contract may involve a positive CCF cost.

To see this, consider the three approaches suggested by Merton (1995) that financial intermediaries could use to protect their customers against intermediary credit risk. The first of these is for the intermediary to match its asset and liability payouts. While many intermediaries do attempt to reduce the maturity gap between their assets and liabilities, they typically do not eliminate this gap. A key reason for this is that maturity transformation is an important economic function served by many intermediaries, so a maturity gap is linked to the *raison d'etre* of the intermediary.

The second approach is for the intermediary to purchase a guarantee from a credible third party. This approach generates a cost, however, when there is moral hazard in the sense that the intermediary can choose a (privately costly) hidden action. In the absence of a guarantee, investors will reflect this moral hazard in the pricing of securities.[28] 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, maturity shortening in response to increased credit risk, and restrictions imposed on the intermediary when covenants are violated.

However, this investor discipline 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 thus exposes customers to it, these customers will flee the intermediary, as discussed in the

[28] 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.

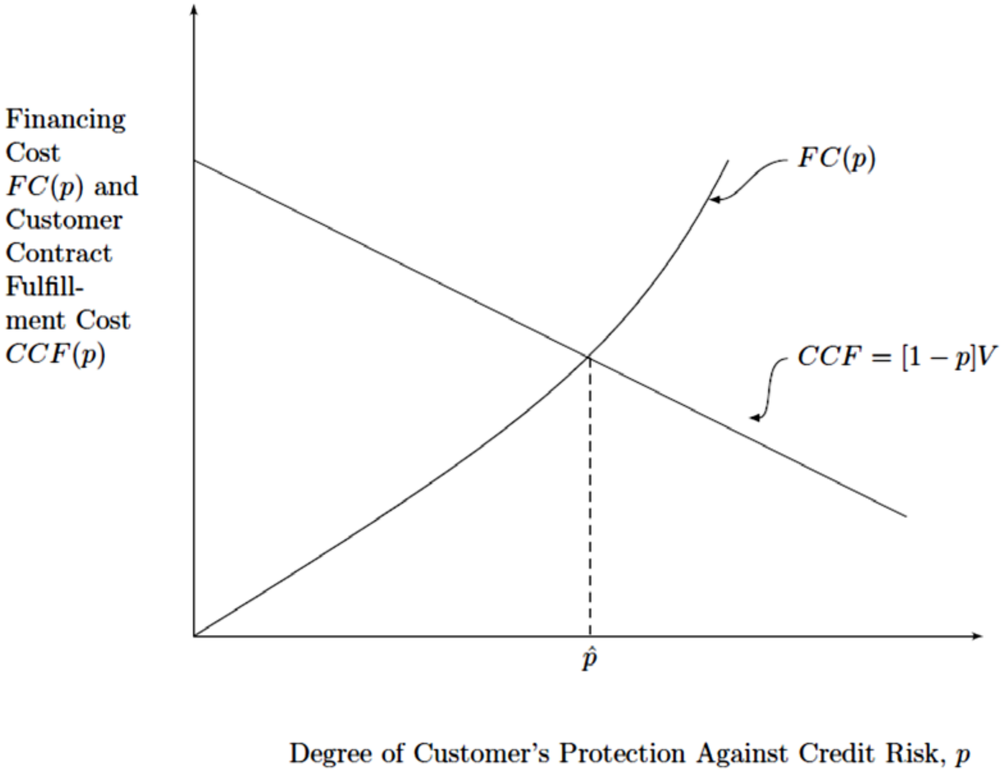
Introduction (see Merton (1997)).[29] This threat of loss of customers can provide additional discipline on the intermediary. By purchasing insurance from a guarantor or by hedging, the intermediary could choose to completely insulate the customer from its own credit risk, but then market discipline from customers would be lost. So the second-best contract may involve customers being (at least potentially) exposed to some credit risk. In the next section, we analyze a formal model that has this feature.

The third approach is for the intermediary to protect its customers by reducing its credit risk through an increase in its equity capital on its balance sheet. However, there is now an extensive literature on the costs that higher equity can entail. This include adverse selection costs (Myers and Majluf (1984)), loss of debt tax shields, loss of debt discipline (e.g. Hart (1995)), and potential loss of liquidity creation (e.g. DeAngelo and Stulz (2015)), among others. These costs may limit the intermediary's reliance on equity capital to reduce credit risk.

In more analytic terms, let $FC(p)$ represent the "financing cost" faced by the intermediary 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 $\partial^2 FC / \partial p^2 \geq 0$, i.e., improving its solvency (and hence reducing the customer's credit risk) is increasingly costly at the margin for the intermediary. 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.

[29] 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.

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



5 An Insurance Application: First-Best and Second-Best Contracts

To fix ideas more precisely, we now develop an application of the concepts in the previous section to insurance. We present both an analysis of the efficient (first-best) contract and the constrained-efficient (second-best) contract. The main difference is that in the first-best case, the intermediary faces no frictions in raising financing, whereas in the second-best case it does.

To summarize the main takeaways, this application:

- Assumes customers are risk averse, so there is a *raison d'etre* for insurance.
- Shows why the customer will prefer to contract with only one insurance company rather than diversify across insurance companies to eliminate idiosyncratic credit risk.
- Shows why in the first-best the insurance company will completely insulate the customer

from its own credit risk.

- Shows why such complete insulation may not be achieved in the second-best, and how this could generate CCF costs.

5.1 The Model

Imagine a two-date world in which there are risk-averse individuals who are purchasing insurance from risk-neutral insurance companies at date $t = 0$. Each individual has initial wealth $w > 0$. There is a probability $\theta \in (0,1)$ that at date $t = 1$, the individual will meet with an accident that will generate a damage of $d \in \mathbb{R}_+$, where $\mathbb{R}_+$ is the positive real line. Let U: $\mathbb{R} \rightarrow \mathbb{R}$ (the real line) be a von-Neumann-Morgenstern utility function over consumption for the individual, with $U(0) = 0$, $U' > 0$, and $U'' < 0$. Assume that all consumption occurs at $t = 1$. The riskfree rate of interest is zero.

We would like to point out that the only purpose of introducing risk aversion in this setting is to create an economic rationale for insurance. Most of the results in this section do *not* rely on customer risk aversion. An alternative that generates a motive for trade and would also work in this context is to have risk-neutral customers who assign a personal disutility to the damage against which they are buying insurance that exceeds the expected cost to the insurance company of providing this insurance. That is, we could stipulate a utility function that is linear in consumption/wealth (unrelated to the accident), $U(w,d) = w - \varphi(d)$, with $\varphi(d) > d \; \forall \; d > 0$, $\varphi(0) = 0$, $\varphi' > 0$, $\varphi'' > 0$, where w is wealth (or consumption) and φ is the customer's disutility of suffering the loss d. All of our results in this section go through with this utility specification.[30]

[30] This specification captures the point we made in Section 3.1, which is that the customer may display risk aversion with respect to the uncertainty about the ability of the intermediary to deliver the embedded promise in the contract itself (in this case to insure the customer against the damage d), but can be risk neutral with respect to all other payoffs.

Suppose the up-front cost of operating as an insurance company is $k \in \mathbb{R}_{+}$. In the context of our earlier analysis (see the discussion preceding Theorem 1), k could be thought of as both the cost of providing insurance services as well as the cost of determining whether the insurance company can profitably provide the services (i.e., whether $\omega \in \Omega_s$).[31] Suppose there are $N > 1$ insurance companies and at $t = 0$, nature randomly assigns each a different cost of providing services k_i, where k_i is drawn from a uniform distribution over $[k_1, k_2, \ldots, k_M]$, with $k_1 < k_2 < \ldots < k_M$, and $M \geq N$. For simplicity, we make an "equal spacing" assumption: $k_1 - k_2 = k_3 - k_2 = \ldots = k_M - k_{M-1}$. Insurance companies act as Bertrand competitors for the individual's insurance business, and must be guaranteed a non-negative profit to participate in the market. Effectively this means that insurance contracts are designed *ex ante* to maximize the customer's expected utility, subject to the participation constraint of the insurance company.

The insurance company has some credit risk. It is captured by the probability $p \in (0,1)$ that the insurance company will be solvent at $t = 1$ and thus able to pay the individual should an insurance claim be filed. Thus, with probability $1 - p$ the customer will be unable to collect on the insurance policy in the event of an accident. We assume (unobservable in the second best) heterogeneity among insurance companies: there is a probability $\delta \in (0,1)$ that the insurer is "strong" financially and a probability $1 - \delta$ that it is "weak". A strong insurer has a probability p_H of being solvent if it chooses effort $e = 1$ and a weak insurer has a probability p_L of being solvent with no ability to change p_L, with $0 < p_L < p_H < 1$. If a strong insurance company chooses $e = 0$, then its probability of solvency becomes p_L. The cost of effort for the insurance company is $e\psi$, where $\psi \in \mathbb{R}_{+}$ is a

[31] One can provide a microfoundation for this by assuming that there are customers who can be profitably insured ($\omega \in \Omega_s$) and those who cannot ($\omega \in \Omega_f$), so a screening or evaluation cost must be incurred by the insurance company to identify the insurable customers. To keep the model simple, we avoid explicitly introducing this customer heterogeneity in the model. A version of the model with this added complexity is available upon request.

positive constant. This effort choice by the insurance company introduces potential moral hazard on the part of the intermediary—this can also be interpreted as investment/underinvestment in services by the intermediary, as discussed previously. If solvent, the insurance company will realize a cash flow of $X \in \mathbb{R}_+$ from which it can make the required payment under the insurance policy. If insolvent, the insurance company's cash flow is zero. The insurance company can purchase a credit default contract from a guarantor that would guarantee the insurance payout to the insured in case the insurance company defaults. The guarantor then plays the role of the investor in this setting.

In analyzing the first-best case, we will assume that strong and weak insurance companies are observably distinct to all, and that a strong company's choice of effort is also costlessly observable. These assumptions will be dropped when we go to the second best. Before we analyze the second best, we will introduce a noisy but informative signal by which customers and investors may be able to infer the intermediary's credit risk. For now, our assumptions imply that this risk is fully observable in the first best.

The sequence of events is described in *Table 1* below.

Table 1: Sequence of Events

***t* = 0**	***t* = 1**
• The value of *k* for each insurance company is realized and publicly observed. • Each insurance company chooses *e* (privately in the second best but publicly in the first best). • Each customer privately observes a noisy but informative signal of *e* (in the second-best case) and decides which insurance company to contract with. It does this before it knows whether the insurance company will be able to purchase a credit default contract from a guarantor, but anticipating the insurance company's equilibrium behavior. • Insurance companies may approach a guarantor to write a credit default contract to guarantee the insurance contract in case the insurance company fails.	• Customer has an accident or does not. • If an accident occurs, the insurance company settles damage claims if it is solvent. Otherwise, the claim is settled by a guarantor if the company entered into the credit default contract. • All consumption occurs.

We will now impose some restrictions on the exogenous parameters to focus on the cases of interest:

$$k_1 - k_2 > \theta d \tag{5}$$

$$[p_H - p_L]X \in \left(\psi, [p_H - p_L]\,\theta d + \psi\right) \tag{6}$$

$$U(w - \theta d - \psi - k_M) > \theta U(w - d) + [1 - \theta]U(w) \tag{7}$$

The inequality in (5) simply states that the cost advantage of the lowest-cost insurer over the second-lowest-cost insurer exceeds the expected payout on the insurance policy. The purpose of this restriction is to generate a high enough equilibrium profit for the lowest-cost insurer so it has sufficient incentive to choose $e = 1$ in the second-best case. This is a sensible restriction since without it the competitive insurance market would never function in a socially efficient manner. The inequality in (6) first says that it is socially efficient for the strong insurer to choose $e = 1$ over $e = 0$, since the increase in the expected payoff due to an increase in the insurer's solvency probability due to the more efficient effort choice exceeds the marginal cost of the effort. Next, (6) says that the strong insurer, left to its own devices, will not choose $e = 1$ in the second-best case if it has an expected payout of θd on the insurance policy; this generates moral hazard and sets the stage for an analysis of second-best distortions. Finally, (7) simply states that insurance has value to the customer even if offered by the highest-possible-cost strong insurance company earning zero expected profit.

5.2 Analysis of the First-Best Contract

We can now prove a series of results that will lead to our central result for this application, namely the complete characterization of the first-best contract between the insurance company and the customer.

Lemma 1: *In the first best, each strong insurance company will choose* $e = 1$. *The customer will sign a contract with the strong insurance company that has the lowest realized k, say* k_l*, and the premium charged by this company will be such that it yields zero expected profit to the insurance company with the second-lowest realized k, say* k_{l2}.

The intuition for this lemma is straightforward. The first part of the lemma follows immediately from (6) and the observability of e. The next part of the lemma follows because, with Bertrand competition, strong insurance companies can always outbid weak insurance companies. Any contract that offers the customer the slightest cost advantage over the best contract offered by the insurance company with the second-lowest cost will lure the customer away to the lowest-cost insurance company. That is, the lowest-cost insurer can win the customer's business with a premium that is epsilon lower than the one that allows the second-lowest-cost insurer to break even. This allows the winning bidder to earn a positive expected profit in the first best.

Lemma 2: *Given that the contract offered is one that enables the insurance company with the second-lowest cost to break even, as shown in Lemma 1, the customer always prefers to pay a higher premium to be completely insulated against the credit risk of the insurance company than to pay a lower premium and be exposed to that risk.*

The customer thus prefers that the insurance company purchase a credit default contract against its own default and include the cost of the contract in the insurance premium, rather than not purchase this protection from a guarantor and expose the customer to its own credit risk. Put together, Lemmas 1 and 2 formalize the intuition that an insurance company that imposes credit risk on its customers will not be able to compete against one that does not.

Lemma 3: *Suppose the customer can completely eliminate its exposure to the credit risk of the insurance company by diversifying across N insurers. In the first best, the customer prefers that the insurance company purchase a credit default contract against its own default on the insurance contract than to diversify on its own across N insurance companies.*

As we alluded to in the previous section, diversifying across many insurance companies is inherently inefficient for the customer because it involves duplicated investigation costs for multiple insurance companies.[32] We can now characterize the first-best insurance contract:

Theorem 3: *The first-best insurance contract involves the customer purchasing insurance from the lowest-cost strong insurer, and this insurer choosing* $e = 1$. *The contract provides complete insurance against the damage caused by the accident, and also completely insulates the customer against the credit risk of the insurer itself. The insurer achieves this insulation by purchasing from a guarantor a credit default contract against its own default on the insurance policy. The premium charged to the customer is:*

$$\theta d + k_{l2} + \psi \tag{8}$$

where k_{l2} *is the investigation cost of the second-lowest-cost insurer.*

Because the customer is risk averse and the insurance company is risk neutral, we know from standard contracting theory that complete insurance against the damage d is optimal. There is no *ex ante* asymmetric information here about the insured, so contracts with incomplete insurance are not offered as part of a separating equilibrium, as in Rothschild and Stiglitz (1976).

There are two noteworthy features of the first-best contract. First, all of the credit risk of the insurance company is transferred to investors (the credit default guarantor in this case), and the customer is willing to pay for investors to bear this risk.[33] Put differently, the customer would not accept a contract that involves exposure to the credit risk of the insurance company in exchange for a lower premium that would leave the insurance company indifferent between the two alternatives. Second, the insurance company offering the contract earns a positive profit in

[32] Note that this result does not depend in any way on customer risk aversion, as is evident in the proof. It will be further reinforced if customers themselves face transaction costs in applying to multiple insurance companies. That is, a risk-aversion-based motive to diversify away idiosyncratic risk is *not* the driving force in Lemma 3.

[33] Note that the insurance premium, $\theta d + k_{l2}$, does not reflect any of the insurer's credit risk, and is higher than the premium, $p_H \theta d + k_{l2}$, that would be charged if the customer were exposed to this credit risk.

equilibrium since its own operating cost is lower than that reflected in the contract.

5.3 Imperfect Observability of Insurer's Credit Risk and the Second-Best

Now we drop the assumption that the insurer's credit risk is perfectly observable to the market. Although an insurance company knows whether it is weak or strong, no one else does—they simply share a common prior belief that the probability that the insurer is strong is δ. We assume instead that there is a noisy signal $\phi_c \in \{p_H, p_L\}$ of the insurance company's credit risk that is first observed privately by each customer for each insurer at $t = 0$ after the (strong) insurer has chosen its effort, e. The distribution of the signal is as follows:

$$\Pr(\phi_c = p_H \mid e = 1, \text{insurer strong}) = q \in (0.5,1) \tag{9}$$

$$\Pr(\phi_c = p_L \mid e = 1, \text{insurer strong}) = 1 - q \tag{10}$$

$$\Pr(\phi_c = p_H \mid e = 0, \text{insurer strong}) = \Pr(\phi_c = p_H, \text{insurer weak}) = 0 \tag{11}$$

$$\Pr(\phi_c = p_L \mid e = 0, \text{insurer strong}) = \Pr(\phi_c = p_L, \text{insurer weak}) = 1 \tag{12}$$

Here, q is the precision of the signal in detecting credit risk. Thus, the signal does not distinguish between a weak insurer (fraction $1 - \delta$ of all insurers) and a strong insurer that has chosen $e = 0$, since both have the same credit risk. Because ϕ_c is privately observed, it is non-verifiable and cannot be used as a conditioning variable in a contract. However, the customer can decide whether or not to enter into a contract with the insurance company after having noisily observed its credit risk, but before knowing whether the insurance company will be able to purchase a credit default contract from a guarantor. Here we are capturing the idea of potential "customer-induced market discipline"—the notion discussed in the Introduction that customers may flee an institution if they are concerned about its ability to fulfill its contractual obligations.

If the customer signs the contract and the insurance company approaches a guarantor for a credit default contract, the guarantor will generate a publicly-verifiable noisy signal $\phi_p \in \{p_H, p_L\}$ of the insurance company's credit risk.[34] This is a signal on which contracts can be conditioned, and we assume for simplicity that this signal and ϕ_c are, conditional on a given e, i.i.d. After observing ϕ_p, the guarantor determines the price and availability of the credit default contract. We now have:

Lemma 4: *Suppose the signals ϕ_c and ϕ_p do not exist. Then it is a unique Nash equilibrium for all (strong) insurance companies to choose $e = 0$ in the second-best case.*

The intuition is that, absent any information about the insurance company's credit risk, neither the insurance premium charged the customer nor the price of the credit default contract can depend on the actual choice of e by the insurance company. In this case, the only benefit of $e = 1$ to the insurance company is the expected enhancement in its net payoff, $X - \theta d$. However, this is insufficient to induce a choice of $e = 1$ when effort is privately costly (see (6)).

The second-best case thus involves an efficiency loss as even strong insurers make effort choices that make them appear as weak insurers. Moral hazard in effort choice eliminates asymmetric information, but in a perverse way as the whole pool of insurers deteriorates in quality. As a consequence, financing costs (related to purchasing credit default contracts) go up for all insurers. The next result shows that this inefficiency persists even if customers privately observe ϕ_c.

Corollary 1: *The result in Lemma 4 holds even if customers observe ϕ_c because customers sign*

[34] One could interpret this signal as a credit rating, or as publicly observable information about credit risk that the rating agency might condition its credit rating on.

contracts regardless of the observed realization of ϕ_c.

The reason for this result is that customers know that the guarantor will be willing to enter into a credit default contract with the insurer, thereby providing complete insurance against damage to the customer. Given this, the customer is indifferent to the credit risk of the insurance company.

Because there is an efficiency gain from lowering the credit risk of the strong insurer that the customer is exposed to, it is useful to explore how the availability of ϕ_p can facilitate the design of a mechanism that can achieve this. In this mechanism, the guarantor pre-commits to offering a credit default contract at a price of $[1 - p_H]\theta d$ to an insurance company for which it observes $\phi_p = p_H$ and denying a contract to an insurance company on which it observes $\phi_p = p_L$. Given this, we assume that insurance companies can sell insurance contracts with a binding precommitment to attempt to purchase a credit default guarantee contract.

Lemma 5: *Given the mechanism above, there exist a precision q high enough and an accident probability θ high enough such that all strong insurers choose e* = 1 *in a Nash equilibrium.*

What induces a strong insurer to choose $e = 1$ now is the fact that by doing so it increases the probability of receiving a credit default contract from zero to q. The customer anticipates this and thus conditions the insurance purchase decision on the observed ϕ_c. That is, unlike the case when the guarantor unconditionally sells credit default contracts to insurance companies, the customer now knows that it will be exposed to the credit risk of the insurer when $\phi_p = p_L$ has been observed, since such an insurer will be unable to purchase credit default protection. The proof of the lemma shows that the customer strictly prefers to be insulated from the credit risk of the insurance company. So insurance is purchased only from an insurance company for which $\phi_p = p_H$. The sufficiency conditions merely require that the customer's signal is precise enough to induce strong reliance on it and the accident probability creates a high demand for insurance.

Our final result is about the second-best contract.

Theorem 4: *Customers purchase insurance contracts only from insurance companies on which* $\phi_c = p_H$ *is observed. The premium charged to the customer is* $\pi(\theta,d) = q\theta d + [1 - q]p_H\theta d + k_{l2} + \psi$, *where* k_{l2} *is the second-lowest cost realization. The second-best insurance contract provides complete insurance against damage d to the insured, but it still exposes the customer to the credit risk of the insurance company with probability* $1 - q$. *The insurance company attempts to purchase credit default protection. If the guarantor observes* $\phi_p = p_H$, *it sells credit default protection; if it observes* $\phi_p = p_L$, *it denies credit default protection.*

The reason why the customer is exposed to the credit risk of the insurance company with a non-zero probability is that, to resolve moral hazard, the guarantor has to deny credit default protection to insurers with high credit risk. However, since the guarantor's signal of the insurer's credit risk is noisy (albeit informative), it sometimes erroneously denies credit default protection to even low-credit-risk (strong) insurers, thereby exposing the customers of that insurance company to credit risk. This is inefficient, but it is a distortion that arises in the second best.[35]

Our more general point is that achieving incentive compatibility and providing incentives for insurance companies to reduce credit risk (and thereby reduce financing frictions) will generally require some consequence for the insurance company that chooses not to invest in credit risk

[35] We do not claim that this mechanism discussed above is the optimal mechanism. For example, it may be more efficient to give insurers with $\phi_p = p_L$ credit default coverage with a non-zero probability, i.e., incentive compatibility for $e = 1$ to be chosen may be achieved at a lower cost. Deriving a unique optimal mechanism will involve additional parametric restrictions, but this additional analysis is unnecessary for the points we want to make.

reduction. What our analysis shows is that this consequence can be created via the customer demand channel, by exposing customers to insurer credit risk should they choose to purchase insurance from the insurers with high credit risks. Doing this may expose even customers who choose to go only with insurers with low credit risks to the insurer's credit risk. This is what happens in our model.[36]

The loss in expected utility for the customer from being exposed to the credit risk of the insurance company can be viewed as a CCF cost that arises in the second-best. Analytically, this cost would be:

$$\begin{aligned}\text{CCF cost} &= [1-q]\Big\{U(w-\pi(\theta,d)) - \{p_H U(w-\pi(\theta,d)) + \big[1-p_H\big][\theta U(w-\pi(\theta,d)-d) \\ &\quad + [1-\theta]U(w-\pi(\theta,d))]\}\Big\} \\ &= [1-q]\theta\big[1-p_H\big]\{U(w-\pi(\theta,d)) - U(w-\pi(\theta,d)-d)\}\end{aligned} \tag{13}$$

Due to the concavity of *U*, it can be shown that the CCF cost as a function of *d* is convex and increasing in *d*.[37]

6. Examples of Customer Contracts, Institutional Design, and Regulatory Practices

In this section, we discuss how our analysis can shed light on some observed financial contracts, institutions, and regulatory practices.

[36] Overall, the main point is that in the second best, exposing customers to *some* of the financial institution's credit risk *may* help to lower financing costs that are elevated by financial frictions, so this may occur despite the inefficiency of doing so relative to the first best.

[37] The CCF cost is positive even with out alternative preference specification of a risk neutral customer with a damage disutility that is increasing and convex in the damage *d*.

6.1 Customer Contracts

6.1.1 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 many ways for a bank to protect its depositors against its own credit risk, although they all entail potential costs. For example, narrow banking, whereby a bank can invest only in safe assets such as U.S. Treasuries, would eliminate bank credit risk but it would require that the bank abandon its key economic services in loan origination, screening, and monitoring; this would represent a potential economic loss. Similarly, requiring that the bank put up a substantial amount of equity may also entail significant costs, depending on the magnitude of the equity infusion, as discussed earlier. Deposit insurance is a solution that avoids those costs, and our analysis offers a rationale for deposit insurance that does *not* rely on preventing runs.[38] Even if contagious bank runs are not a problem, deposit insurance improves efficiency in our theory because it enables one to move closer to the first best in which the bank's customers are completely insulated from its credit risk. Deposit insurance is one reason why customers (retail depositors) are willing to deal with institutions that lack a AAA credit rating.[39] The fact that depositor insurance is incomplete

[38] Preventing runs is the most widespread justification for deposit insurance in the literature, dating back to Bryant (1980) and Diamond and Dybvig (1983).

[39] These ratings refer to the credit risk to which the bank's uninsured creditors (investors) are exposed, not its depositors. Deposit insurance makes the deposit contract effectively riskfree for the (core) depositors—those with deposit balances below the coverage limit—even if the bank lacks a AAA rating on its uninsured debt. These core depositors represent the bulk of the customers our framework focuses on, with perhaps a small fraction of these customers being partially insured and representing a source of market discipline.

may be understood in the context of our analysis of the second-best insurance contract, namely the need for customer-imposed market discipline to deal with moral hazard.[40]

6.1.2 Mutual Funds

For mutual funds, customers are investors in the fund—each customer is purchasing a service (the portfolio management service and the promise of some risky return), while also providing financing. In this case, 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—say, due to unobserved risky investments with fund money or "tunneling"—that the customer wishes to be insulated from.

A mutual fund is a good example of a contract that imposes risk on the investor (customer), but only that risk which is confined to the contract itself.[41] Indeed, this is one reason why investors put their money in funds managed by reputable intermediaries like Vanguard, Fidelity, and the like. If one invests in the S&P 500 through one of these funds, the risk (*R*) is $R_{S\&P\ 500}$. 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.

6.1.3 Insurance Contracts

An individual who purchases a whole life insurance policy is a customer who is buying a

[40] Merton (1977) shows how deposit insurance is isomorphic to a put option, and how it can create moral hazard.

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

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 a risk-sharing distortion because it would not be efficient for the insured to buy a large number of smaller life insurance policies to diversify across insurance companies, as we showed earlier.

A similar argument holds for property and casualty insurance. Our analysis in the previous section shows why the first-best insurance contract completely protects the customer from the credit risk of the insurance company. To the extent that the second best may leave the customer with some exposure, an insurance fund that backs up insurance companies and protects policyholders can enhance efficiency. In all fifty U.S. states, state insurance funds provide this service.[42] This moves the second-best contract closer to the first best.

6.1.4 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 theory. 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

[42] 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.

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 intermediary. As a result, the second-best arrangement approximates the efficiency of the first best.[43] In fact, the efficiency of the first best is achieved by the central bank repo facility (known as RRP) started by the Federal Reserve in 2013. This is an overnight repo program in which institutions are able to park cash with the Fed and earn a modest interest rate in exchange for U.S. Treasuries as collateral. Our theory helps to explain the growing popularity of this facility.[44]

6.2 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.[45] 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

[43] Indeed, whenever there are concerns 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.

[44] See Burne (2016).

[45] 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 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.

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.

6.3 Regulatory Practices: 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 analysis provides an economic rationale for this. By making swaps exchange-traded, counterparty credit risk is greatly reduced, moving the arrangement closer to first best. 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. Thus, this requirement of Dodd-Frank serves the economic purpose of minimizing customer-specific contract fulfillment risk in swaps.

An interesting question is why market participants did not do this on their own prior to Dodd-Frank. There are a variety of possible explanations for this. One reason may be coordination failures among market participants in the process of setting up an exchange. The fact that coordination failures can lead to adoption externalities that cause individual participants to avoid welfare-enhancing initiatives has been established in various contexts.[46] For example, Dybvig and

[46] Coordination failures in our context can take many forms. For example, low-default-risk counterparties who are privately informed about each other’s credit risk may prefer to engage in off-exchange bilateral contracts and avoid exchange-trading costs. This may result in a downward quality spiral in firms that voluntarily choose exchange trading, a problem that is potentially exacerbated by higher costs of setting up an exchange, which may be the case with highly-customized swap contracts. This can adversely affect liquidity in exchange trading.

Spatt (1983) develop a model which explains that the failure to adopt the metric system in the U.S. may be attributable to such a coordination failure. Fishman and Hagerty (2003) develop a model in which mandatory information disclosure is rationalized on the grounds that voluntary disclosure may not be forthcoming when the fraction of consumers who can understand a disclosure is too low. These are examples of economic settings in which a regulatory mandate helps to overcome the negative adoption externalities generated by coordination failures.

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 to bear no intermediary credit risk. 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 in our framework 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 intermediary-specific 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, and repos. It also provides a fresh perspective on why deposit insurance may be efficiency-enhancing even in the absence of contagious runs, and why we have securities exchanges. Moreover, it generates an economic rationale for the swaps clearinghouse requirement of the Dodd-Frank Act. An empirical implication of our theory is that whenever customers' concerns about the credit risks of the financial institutions they deal with are elevated, those institutions that offer their customers better protection will gain a competitive advantage.

We view this theoretical framework as a useful starting point for identifying and understanding how the key roles of customers and investors impact financial intermediaries. Future work could take the framework further, and examine the implications for regulation policy involving systemic risks. In addition, our theory 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 approach bears on the work related to opacity and transparency in contracts.

References

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

2) Boot, Arnoud WA. "Relationship Banking: What Do We Know?." *Journal of Financial Intermediation* 9, no. 1 (2000): 7-25.

3) Boot, Arnoud WA, and Anjan V. Thakor. "Security design." *The Journal of Finance* 48, no. 4 (1993): 1349-1378.

4) Bryant, John. "A model of reserves, bank runs, and deposit insurance." *Journal of Banking & Finance* 4, no. 4 (1980): 335-344.

5) Burne, Katy, “Fed Repo Program Swells”, *The Wall Street Journal*, October 3, 2016. p. C7.

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

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

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

9) DeAngelo, Harry, and René M. Stulz. "Liquid-claim production, risk management, and bank capital structure: Why high leverage is optimal for banks." *Journal of Financial Economics* 116, no. 2 (2015): 219-236.

10) DeMarzo, Peter M. "The pooling and tranching of securities: A model of informed intermediation." *Review of Financial Studies* 18, no. 1 (2005): 1-35.

11) DeMarzo, Peter, and Darrell Duffie. "A liquidity-based model of security design." *Econometrica* 67, no. 1 (1999): 65-99.

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

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

14) Dybvig, Philip H., and Chester S. Spatt. "Adoption externalities as public goods." *Journal of Public Economics* 20, no. 2 (1983): 231-247.

15) Erel, Isil, Taylor Nadauld, and René M. Stulz. "Why Did Holdings of Highly Rated

Securitization Tranches Differ So Much across Banks?" *Review of Financial Studies* 27, no. 2 (2014): 404-453.

16) Fishman, Michael J., and Kathleen M. Hagerty. "Mandatory versus voluntary disclosure in markets with informed and uninformed customers." *Journal of Law, Economics, and organization* 19, no. 1 (2003): 45-63.

17) 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.

18) Fulghieri, Paolo, and Dmitry Lukin. "Information production, dilution costs, and optimal security design." *Journal of Financial Economics* 61, no. 1 (2001): 3-42.

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

20) Hanson, Samuel G., Andrei Shleifer, Jeremy C. Stein, and Robert W. Vishny. "Banks as patient fixed-income investors." *Journal of Financial Economics* 117, no. 3 (2015): 449-469.

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

22) Hart, Oliver, and Luigi Zingales. “Banks Are Where The Liquidity Is.” No. w20207. National Bureau of Economic Research, 2014.

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

24) 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.

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

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

27) Krishnamurthy, Arvind, and Annette Vissing-Jorgensen. "The aggregate demand for treasury debt." *Journal of Political Economy* 120, no. 2 (2012): 233-267.

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

29) 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.

30) 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.

31) Merton, Robert C. "The financial system and economic performance." In *International Competitiveness in Financial Services*, pp. 5-42. Springer Netherlands, 1990.

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

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

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

35) 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.

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

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

38) 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 et. al., 3–31. Boston: Harvard Business School Press, 1995.

39) 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.

40) Myers, Stewart C., and Nicholas S. Majluf. "Corporate financing and investment decisions when firms have information that investors do not have." *Journal of Financial Economics* 13, no. 2 (1984): 187-221.

41) 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.

42) Rothschild, Michael, and Joseph Stiglitz. "Equilibrium in competitive insurance markets: An

essay on the economics of imperfect information." In *Foundations of Insurance Economics*, pp. 355-375. Springer Netherlands, 1976.

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

Appendix

Proof of Theorem 1: Note that $\frac{\partial ES_{total}}{\partial p} = \frac{\partial ES_c}{\partial p} = V$. Moreover, $\frac{\partial^2 ES_{total}}{\partial p \partial V} = \frac{\partial^2 ES_c}{\partial p \partial V} > 0$. Now p^0 is the value of p that satisfies $ES_c = 0$, so $p^0 = \overline{V}/V$. Similarly, p^* is the value of p that satisfies $ES_{total} = 0$. Thus, $p^* = \left[\overline{V} + k - V_m^I\right]/V$. This yields:

$$1 - p^0 = [V - \overline{V}][V]^{-1} \qquad \text{(A-1)}$$

$$1 - p^* = [V - \overline{V} + V_m^I - k][V]^{-1} \qquad \text{(A-2)}$$

It follows from (1) and (2) that $1 - p^0 \geq 0$ and $1 - p^* > 1 - p^0$. Finally, note that $\frac{\partial p^*}{\partial [V_m^I - k]} = -\frac{1}{V} < 0$.
■

Proof of Theorem 2: The proof is straightforward for the case in which the customer is risk averse. When the customer is risk neutral, the expected value of the service provided by the intermediary is $\Pr(\omega \in \Omega_s)pV + [1 - p]\lambda\overline{V}$, where λ is the probability that another institution can provide the service at $t = T$ in case the institution that contracted with the customer at $t = 0$ fails. Since the replacement intermediary cannot generate a value higher than the reservation value $\overline{V}$, that is the term that appears above. Note that if the institution is solvent, it only provides the service when its information about the customer reveals $\omega \in \Omega_s$, and this holds for all institutions providing the service.

Now note that the unprofitability of providing the service when $\omega \in \Omega$ means that $\lambda = 0$ if k_c is high enough. Even if $k_c \in (k,\infty]$ is not prohibitive, $\overline{V} < V$ since $k_c > k$. Hence, the social surplus generated by the customer's relationship with the institution is $\Pr(\omega \in \Omega_s)pV$, which is maximized at $p = 1$. ■

Proof of Lemma 1: The social gain from choosing $e = 1$ relative to $e = 0$ is $\left[p_H - p_L\right]X$, and the social cost is ψ. By (6), the social gain from choosing $e = 1$ strictly exceeds the cost. This proves the first part of the lemma. For the next part of the lemma, since each of the N insurance companies has a different realized k_i, it must be the case that one can rank-order the realized k's across the N insurance companies, $\{k_{\mathrm{l}}, \dots ,k_N\} \subseteq \{k_1, \dots ,k_M\}$. Let k_l be the lowest k and k_{l2} be the second lowest k. Then with Bertrand competition among the insurance companies, the strong insurance company i with $k_i = k_l$ can offer a contract that charges a premium of $\pi(\theta,d) - \varepsilon$, where

$$\pi(\theta,d) = p_H\theta d + \psi + k_{l2} \tag{A-3}$$

and $\varepsilon > 0$ is arbitrarily small such that the insurance company with $k_i = k_{l2}$ earns zero expected profit with a premium of $\pi(\theta,d) + k_{l2}$. The insurance company with $k_i = k_{l2}$ cannot match this offer because it violates its participation constraint, whereas the insurance company with $k_i = k_l$ earns a positive expected profit since $k_l < k_{l2}$. The proof is completed by letting $\varepsilon \to 0$. ■

Proof of Lemma 2: If the customer is exposed to the credit risk of the insurance company, the premium charged is

$$p_H\theta d + k_{l2} + \psi \tag{A-4}$$

and the customer is uninsured with probability $1 - p_H$. If the insurance company buys a credit default contract against its own credit risk, then the premium will be

$$\theta d + k_{l2} + \psi \tag{A-5}$$

and the customer is completely insured regardless of the insurance company's credit risk. Now, the customer's expected utility when exposed to the credit risk of the insurance company is:

$$\begin{aligned}
U_0 &= p_H U(w - p_H\theta d - k_{l2} - \psi) \\
&\quad + [1 - p_H]\{\theta U(w - p_H\theta d - d - k_{l2} - \psi) + [1 - \theta]U(w - p_H\theta d - k_{l2} - \psi)\} \\
&< p_H U(w - p_H\theta d - k_{l2} - \psi) + [1 - p_H]\{U(\theta w - \theta p_H\theta d - \theta d - \theta k_{l2} - \theta\psi \\
&\quad + [1 - \theta]w - [1 - \theta]p_H\theta d - [1 - \theta]k_{l2} - [1 - \theta]\psi)\} \\
&\text{(by Jensen's inequality)} \\
&= p_H U(w - p_H\theta d - k_{l2} - \psi) + [1 - p_H]U(w - p_H\theta d - \theta d - k_{l2} - \psi) \\
&< U(p_H w - p_H^2\theta d - p_H k_{l2} - p_H\psi + [1 - p_H]w - [1 - p_H]p_H\theta d \\
&\quad - [1 - p_H]\theta d - [1 - p_H]k_{l2} - [1 - p_H]\psi) \\
&\text{(by Jensen's inequality)} \\
&= U(w - p_H\theta d - [1 - p_H]\theta d - k_{l2} - \psi) \\
&= \widehat{U}_0
\end{aligned} \tag{A-6}$$

where $\widehat{U}_0$ is the customer's expected utility when completely protected against the credit risk of the insurance company. ■

Proof of Lemma 3: Suppose the customer can completely eliminate exposure to the credit risk of the insurance company by diversifying across *N* insurance companies, none of which purchases a credit default contract, but all of which incur the cost *k*. Then, assuming that each insurance company can spread its effort cost ψ over *N* customers, the customer's expected utility is:

$$\begin{aligned} U_0 &= U\left(w - \theta d - \sum_{i=2}^{N} k_{li} - \hat{k}_{l2} - \psi \right) \\ &< U\left(w - \theta d - \hat{k}_{l2} - \psi\right) \\ &= \hat{U}_0 \end{aligned} \tag{A-7}$$

where $\hat{k}_{l2}$ is the realized investigation cost of the second-lowest-cost insurer (which is reflected in the price of the contract offered by the lowest-cost insurer), and the summation $\sum_{i=2}^{N} k_{li}$ is from the second-lowest-cost insurer to the highest-cost insurer, and $\hat{U}_0$ is the expected utility of the customer when going with only the lowest-cost insurer who has purchased a credit default contract against its own default on the insurance contract. ■

Proof of Theorem 3: From Theorem 1, we know that the first-best contract completely protects the customer against the credit risk of the insurance company. Moreover, because the customer is risk averse and the insurance company is risk neutral, the contract must also offer the customer complete protection against the damage caused by the accident. Given the feasible contract space, it follows that the insurance company will purchase a credit default contract from a guarantor to insulate the customer from the insurance company's credit risk. The expression for the insurance premium in (8) follows from the Bertrand competition condition whereby the insurer prices the policy to generate zero profit for the insurer with the second-lowest cost $k_{\mathrm{l}2}$. ■

Proof of Lemma 4: Let ζ be the price of the credit default contract. For the lemma to be true, we need:

$$\psi + \theta d + k_{l2} - k_l + p_L[X - \theta d] - \zeta > \theta d + k_{l2} + \psi - k_l + p_H[X - \theta d] - \zeta - \psi \tag{A-8}$$

where the left-hand side is the insurance company's expected payoff if it chooses $e = 0$ and the right-hand side is its expected payoff if it chooses $e = 1$. Here k_l is the lowest-cost realization and k_{l2} is the second-lowest cost realization. Rearranging (A-8) gives:

$$\left[p_H - p_L\right][X - \theta d] < \psi \tag{A-9}$$

which holds, given (6). ■

Proof of Corollary 1: The customer's expected utility is $U(w - \theta d - \psi - k_{l2})$ regardless of the realization of ϕ_c if the insurance company always succeeds in buying a credit default contract. The guarantor can break even by pricing the contract at $\left[1 - p_L\right]\theta d$, so it will be willing to offer the contract even though all insurers have $p = p_L$. ■

Proof of Lemma 5: Note that the customer's posterior belief after observing ϕ_c is:

$$\begin{aligned}&\text{Pr(insurer can purchase a credit default contract} \mid \phi_c)\\ &= \text{Pr(insurer strong and chosen } e = 1 \mid \phi_c) \times \Pr(\phi_p = p_H \mid e = 1, \text{insurer strong}) \qquad \text{(A-10)}\end{aligned}$$

First, we establish that a customer who observes $\phi_c = p_H$ will enter into a contract with the insurance company, and one who observes $\phi_c = p_L$ will not. Now

$$\text{Pr(insurer is strong and has chosen } e = 1 \mid \phi_c = p_H) = 1 \qquad \text{(A-11)}$$

Thus, since $\Pr(\phi_p = p_H \mid e = 1, \text{insurer strong}) = q$, the customer assesses the probability that the insurer will be able to purchase a credit default contract as q. This implies that the probability of the customer being completely protected against the credit risk of the insurance company is q. Hence, the customer's expected utility is:

$$U_H = qU(w - \pi) + [1 - q]\{\theta[p_H U(w - \pi) + [1 - p_H]U(w - \pi - d)] + [1 - \theta]U(w - \pi)\} \qquad \text{(A-12)}$$

where π is the insurance premium when $\phi_c = p_H$. If the customer observes $\phi_c = p_L$, then

$$\text{Pr(insurer is strong and has chosen } e = 1 \mid \phi_c = p_L) = \frac{[1 - q]\delta}{[1 - q]\delta + 1 - \delta} \qquad \text{(A-13)}$$

given the Nash equilibrium strategy of all strong insurers to choose $e = 1$. If the customer goes with such an insurance company, the customer's expected utility will be:

$$\begin{aligned}U_L = &\left(\frac{[1 - q]\delta}{[1 - q]\delta + 1 - \delta}\right) qU(w - \tilde{\pi}) + \left(\frac{[1 - \delta] + [1 - q]^2\delta}{[1 - q]\delta + 1 - \delta}\right)\{\theta[p_H U(w - \tilde{\pi})\\ &+ [1 - p_H]U(w - \tilde{\pi} - d)] + [1 - \theta]U(w - \tilde{\pi})\} \qquad \text{(A-14)}\end{aligned}$$

where $\tilde{\pi}$ is the insurance premium associated with the customer observing $\phi_c = p_L$.

Now consider a scenario where the customer has observed $\phi_c = p_L$ for the insurance company that has realized $k = k_1$ and has observed $\phi_c = p_H$ for the insurance company that has realized $k = k_M$. Assume each offers the customer a contract that yields the insurance company zero expected profit. We will show that for q high enough, the customer's expected utility is higher in going with the insurance company for which $\phi_c = p_H$ has been observed. That is, we want to show:

$$U_H > U_L \tag{A-15}$$

Note that the competitive breakeven premium pricing condition implies that if $\phi_c = p_H$ for the insurer with $k = k_M$ and $\phi_c = p_L$ for the insurer with $k = k_1$, then:

$$\pi = q\theta d + [1 - q]p_H\theta d + \psi + k_M \tag{A-16}$$

Similarly,

$$\tilde{\pi} = \hat{q}\theta d + [1 - \hat{q}]p_H\theta d + \psi + k_1 \tag{A-17}$$

where

$$\hat{q} \equiv \frac{[1 - q]\delta q}{[1 - q]\delta + [1 - \delta]} < q \tag{A-18}$$

Now, we will show that $U_H > U_L$ for q and θ high enough.

Note that (A-16) and (A-17) tell us that

$$\lim_{q \to 1} \pi = \theta d + \psi + k_M \equiv \hat{\pi}_M \tag{A-19}$$

$$\lim_{q \to 1} \tilde{\pi} = p_H\theta d + \psi + k_1 \equiv \hat{\pi}_1 \tag{A-20}$$

Moreover, at $q = 1$, we have:

$$U_L(q = 1) = \theta[p_H U(w - \hat{\pi}_1) + [1 - p_H]U(w - \hat{\pi}_1 - d)] + [1 - \theta]U(w - \hat{\pi}_1)$$

$$< \theta U\big(p_H w - p_H\hat{\pi}_1 + [1 - p_H]w - [1 - p_H]\hat{\pi}_1 - [1 - p_H]d\big) + [1 - \theta]U(w - \hat{\pi}_1)$$

(by Jensen's inequality)

$$= \theta U\big(w - \hat{\pi}_1 - [1 - p_H]d\big) + [1 - \theta]U(w - \hat{\pi}_1)$$

$$= \theta U\big(w - d - \psi - k_1 + p_H d[1 - \theta]\big) + [1 - \theta]U(w - \hat{\pi}_1)$$

$$< \theta U(w - d\,) + [1 - \theta]U(w)$$

(since $\psi + k_1 > p_H d[1 - \theta]$ for θ high enough)

$$< U(w - \theta d - \psi - k_M)$$

(by (7))

$$= U_H(q = 1) \tag{A-21}$$

This means that the customer strictly prefers to purchase insurance from a company with $\phi_c = p_H$ than from a company with $\phi_c = p_L$, regardless of their k realizations. Moreover, the result that $U_L(q = 1) < \theta U(w - d\,) + [1 - \theta]U(w)$ means that the customer would prefer to be uninsured than to purchase insurance from a company with $\phi_c = p_L$. Thus, an insurance company that has $\phi_c = p_L$ observed by the customer can never win the customer's business, regardless of its realized *k*.

Consider now the problem of the insurance firm that has realized the lowest *k*. The IC constraint for it to prefer choosing *e* = 1 over *e* = 0 is:

$$q\left\{k_{l2} - k_l + q\theta d + [1-q]p_H\theta d + \psi + p_H[X - \theta d] - q\left[1 - p_H\right]\theta d\right\} + [1-q]p_H X - \psi$$
$$\geq p_L X \tag{A-22}$$

where k_l is the investigation cost of the lowest-cost insurer and k_{l2} is the cost of the second-lowest-cost insurer. To understand (A-22), note that the left-hand-side has three terms. The first term has *q*, the probability the customer will observe $\phi_c = p_H$ given *e* = 1, which multiplies the term in braces, which is the expected profit of the insurer if it gets the insurance contract. It includes the premium $k_{l2} + q\theta d + [1-q]p_H\theta d + \psi$, which is determined this way because the customer computes a probability $\Pr\left(\phi_p = p_H\right) = q$ that the guarantor will observe $\phi_p = p_H$ and give a credit default contract to the insurer so that the customer is fully covered, and the probability is $1 - q$ that the insurer will not get a credit default contract so the customer will be covered only when the insurer is solvent (probability p_H). It also includes the net expected payoff to the insurance company when the company is solvent and incurs its expected obligation of θd under its insurance contract, $X - \theta d$, multiplied with the probability of solvency, p_H. Subtracted from this is the insurance company's investigation cost k_l and the expected credit default contract price $q\left[1 - p_H\right]\theta d$, which is the credit default price $\left[1 - p_H\right]\theta d$ the insurance company pays with probability *q* (the probability that the guarantor observes $\phi_p = p_H$). With probability $1 - q$, the customer observes $\phi_c = p_L$, so the insurance company does not get the customer's contract. As a result, its expected payoff is $p_H X$. Thus, the second term is $[1-q]p_H X$. Finally, the third term is the insurance company's cost of effort ψ. The right-hand-side is the insurance company's expected payoff from choosing *e* = 0. With probability one, $\phi_c = p_L$ is observed and no contract is awarded, so the expected payoff is $p_L X$.

Simplifying (A-22) yields:

$$q\left\{k_{l2} - k_l + p_H X\right\} + [1-q]p_H X - [1-q]\psi \geq p_L X \tag{A-23}$$

which becomes:

$$q\left\{k_{l2}-k_{l}\right\}+X[p_{H}-p_{L}]\geq[1-q]\psi \qquad \text{(A-24)}$$

which clearly holds given (6) and the fact that $k_{l2} > k_l$. ■

Proof of Theorem 4: In the proof of Lemma 5, we already showed that customers purchase insurance contracts only when they observe $\phi_c = p_H$. We also showed that the scheme of the guarantor selling a credit default contract to the insurer only if $\phi_p = p_H$ is observed generates a probability $1 - q$ that even a strong insurer that chooses $e = 1$ will not get a credit default contract. ■