Bank Screening and Monitoring*

Sebastian Gryglewicz† Simon Mayer‡ Erwan Morellec§

August 3, 2022

Abstract

We develop a dynamic agency model in which a bank originates loans that it can sell to investors. The bank controls the default risk of the loans through screening at origination and monitoring after origination, but is subject to moral hazard. The optimal contract between the bank and investors can be implemented via a time-decreasing stake in the loan, rationalizing loan sales after origination. Loan characteristics affect loan performance and initial retention and shape selloff dynamics. Screening and monitoring have positive incentive synergies, making it optimal to task loan originators with monitoring. Credit ratings distort incentives, potentially increasing credit risk.

---

*We would like to thank Doug Diamond, Andreas Fuster, Thomas Geelen, Denis Gromb, Barney Hartman-Glaser, Shohini Kundu, Gustavo Manso, Ralf Meisenzahl, Martin Oehmke, Alejandro Rivera, Martin Schmalz, Vish Viswanathan, and seminar participants at the MFA 2022, and the 2022 FTG meeting in Budapest for comments. Erwan Morellec acknowledges financial support from the Swiss Finance Institute.

†Erasmus University Rotterdam. Email: gryglewicz@ese.eur.nl
‡HEC Paris. E-mail: simon.r.mayer@gmail.com.
§EPF Lausanne, Swiss Finance Institute, and CEPR. E-mail: erwan.morellec@epfl.ch.
Over the past 20 years, outstanding corporate debt in the U.S. has nearly tripled. This increase has been fueled by the emergence of an active and liquid secondary market for corporate loans, where loans are traded like debt securities (Saunders, Spina, Steffen, and Streitz (2021)), as well as by the growth of collateralized loan obligations (CLOs), in which a broad array of nonbank financial institutions invest (Cordell, Roberts, and Schwert (2021)). These developments have given loan originators the possibility to reduce or eliminate their exposure to borrowers’ default risk by selling their stake over the loan’s life (Blickle, Fleckenstein, Hillenbrand, and Saunders (2022)). As a result, concerns have been expressed that problems in the corporate debt markets are building up in a similar way as they did in the run-up to the subprime mortgage market crisis.

A key difference between mortgages and corporate loans is that in addition to the screening that takes place prior to issuance, a bank reduces risk and adds value to corporate loans through frequent monitoring over the life of the loan. However, if the bank sells (part of) the loans it has originated, it may not have sufficient incentives to screen and monitor borrowers (Pennacchi (1988) or Gorton and Pennacchi (1995)). While securitization and its consequences for the mortgage market have been the subject of considerable empirical and theoretical research, much less is known about the relation between skin in the game (i.e. the share retained by loan originators) and screening, monitoring, and default risk in corporate loan markets. In this paper, we develop a tractable, dynamic framework to study optimal incentive provision for screening and monitoring in the context of corporate loans. We then derive implications for (dynamically) optimal originator share, the value of credit ratings, and their effects on credit risk. This allows us to (i) examine whether observed arrangements are rational, (ii) shed light on existing empirical findings, and (iii) generate new implications regarding the effects of loan characteristics such as maturity on screening, monitoring, and default risk.

We start our analysis by formulating a dynamic agency model in which a bank (the agent) originates a loan and sells this loan to competitive investors (the principal; e.g., other financial institutions in a loan syndicate). The loan generates coupon payments at a constant rate up to default or maturity. When originating the loan, the bank may undertake a costly screening effort that results in a lower expected default rate. It can also continuously monitor the loan at a cost

1The majority of loans traded in the secondary market are syndicated loans, i.e., loans issued to a borrower jointly by multiple financial institutions under one contract. The syndicated loan market is one of the most important sources of private debt for corporations with an annual primary market issuance volume in the U.S. that exceeded that of public debt and equity as early as 2005 (Sufi (2007)).

2CLOs operate as special purpose vehicles that issue tranched asset-backed securities or notes to investors, and use the proceeds to finance the purchase of leveraged loans. See Kundu (2021) for an in-depth analysis of CLOs.

3As documented for instance in Benmelech, Dlugosz, and Ivashina (2012), the securitization of corporate loans is fundamentally different from the securitization of other asset classes. Corporate loans are significantly larger than mortgages and are typically syndicated. The bank that originated the loan generally retains a fraction of the loan on its balance sheet. Fractions of the same underlying loan are simultaneously held by CLOs as well as by other institutional investors and banks. In addition, each loan included in CLOs is rated.
afterward to further control default risk. The loan default intensity is endogenous in that the loan will have a higher default rate if the bank decides to shirk. Screening at origination and monitoring after origination are not observable by investors, leading to moral hazard. The bank has a lower valuation for the loan than investors due to a higher discount rate arising, e.g., from regulatory or capital constraints. There are therefore gains from selling (part of) the loan to investors. Loan sales reduce the bank’s exposure to loan performance and undermine its incentives, thereby increasing credit risk and reducing the loan value.

We derive the optimal contract between the loan originator and outside investors that implements costly screening and monitoring. We do not impose any restriction on the form of the contract and include all possible payment schedules, so long as they provide limited liability to both the bank and investors. Incentive provision for screening and monitoring requires exposing the bank to loan performance. As the bank is protected by limited liability, this is achieved by delaying its payouts so that the bank loses its expected future payments upon default. However, delaying payments is costly due to the bank’s higher discount rate. Based on this trade-off, the paper derives an incentive compatible contract that maximizes total surplus. This contract takes a simple form: The bank retains a share of the loan at origination and gradually sells the loan over time, in line with observed practice.

The structure of the optimal contract arises from positive spillovers between screening and monitoring. Notably, the exposure to loan performance that is necessary to provide monitoring incentives after origination generates additional screening incentives at origination (by increasing the agent’s skin in the game), leading to synergies between screening and monitoring. These synergies also imply that the optimal contract provides high monitoring incentives due to moral hazard over screening. As screening only occurs at origination, the optimal contract front-loads incentives, so the agent’s incentives by means of delayed payouts are especially strong at origination and decrease over time. Accordingly, monitoring incentives decrease while default risk increases over time. To achieve this reduction in deferred compensation and monitoring incentives, the optimal contract mandates smooth, time-decreasing payments to the agent. Therefore, the optimal contract can be implemented by requiring the loan originator (the lead bank in the case of syndicated loans) to retain a stake in the loan that it gradually sells to investors.

In our model, both screening and monitoring increase with the bank’s incentives—i.e. with its stake in the loan—and are negatively associated with loan spreads, in line with the evidence on screening in Ivashina (2009) and on monitoring in Wang and Xia (2014) and Gustafson, Ivanov, and Meisenzahl (2021). What determines the initial level of retention and the optimal speed of loan sales? The model predicts that the originating bank initially retains a significant fraction of
the loan, in line with the evidence in Benmelech et al. (2012) and Gustafson et al. (2021). Initial retention is lower when intrinsic (i.e., pre-screening) credit risk is high, which is the case when the loans are of bad quality or their underlying collateral is risky, when moral hazard is severe due to high screening or monitoring costs (due to, e.g., a higher fraction of soft information), when loan maturity is long, or when the originator’s capital constraints are large. A unique feature of our model is that it predicts that the originator’s (or lead bank’s) share in the loan should decrease through time, in line with the recent evidence in Blickle et al. (2022). In our model, loan sales after origination arise as part of the optimal contract, which implies that incentives are front-loaded. We show that loan sales decrease lenders’ incentives to monitor, thereby increasing the default intensity and hence credit risk. Our paper additionally predicts that the speed of loan sales should decrease with the loan quality at origination, as captured by intrinsic loan quality, and should increase with the cost of monitoring and the bank’s cost of capital.

Interestingly, we also show that screening and monitoring are complements, in that an increase in the cost of screening or monitoring leads to a decrease in the optimal levels of both screening and monitoring. The reason is that when, for instance, monitoring is costly, it is optimal to reduce monitoring incentives. As screening and monitoring incentives exhibit synergies, the reduction in monitoring incentives reduces screening incentives. Our paper additionally shows that a decrease in intrinsic borrower quality reduces the marginal impact of screening and monitoring and leads to laxer monitoring and screening, thereby further exacerbating credit risk. Through this mechanism, our model provides a rationale for the segmentation observed in credit markets. According to our analysis, banks that exert high screening and high monitoring typically finance high quality borrowers with high priority loans. By contrast, private equity and online lenders finance lower quality borrowers with low priority debt instruments. Our analysis also suggests that when screening is more lax, monitoring should also be more lax. It is therefore consistent with the trend observed in the leveraged loan market (i.e., the riskier segment of the market), in which the incidence of including covenants is decreasing and where more than 80% of outstanding loans in 2020 were covenant light according to Standard & Poor’s.

One way for loan originators to reduce their skin in the game is to use securitization, for example by including CLOs in the syndicate. As discussed in Daley, Green, and Vanasco (2020), the development of markets for securitized products has been facilitated in part by credit rating agencies, “which allow issuers access to a large pool of investors who would otherwise have perceived these securities as opaque and complex.” Indeed, a feature that CLOs share is that each loan

---

4 Relatedly, Ivashina and Vallée (2021) find in recent research that weakening clauses in loan contracts (i.e., clauses that weaken covenants) are particularly common when banks retain a smaller share of the loan.
included in the deal gets rated. By providing information about initial credit quality, credit ratings at origination generate screening incentives, as lax screening induces a low rating. In the aftermath of the 2008 financial crisis, during which highly rated structured debt products performed poorly, the roles of originators in screening loans and of rating agencies in evaluating securitized products have come into question. According to some, this poor performance suggested that “the initial ratings of structured debt securities greatly understated their risk” (Pagano and Volpin (2010)). While credit ratings generate incentives to screen, they leave moral hazard over monitoring unchanged. Accordingly, our model implies that a credit rating at origination induces not only more screening but also lower incentives for the agent through delayed payments. That is, our model predicts that initial retention of the lead bank is lower when CLOs invest. Therefore, credit ratings weaken monitoring incentives—i.e. the model predicts that monitoring should be less intensive for syndicated loans with CLOs—and, as a result, have an ambiguous impact on credit risk.

In some applications of credit securitization, screening and monitoring of loans are undertaken by separate entities: an originator responsible for screening and a servicing company in charge of monitoring (Demiroglu and James (2012)). In other settings, monitoring and screening are undertaken by the same entity. An important question is therefore whether bundling affects incentives and credit risk. To answer this question, we consider a model variant in which two otherwise identical agents, respectively called screener and monitor, respectively screen and monitor loans and are both subject to moral hazard. For the screener and monitor to have adequate incentives, they must retain a stake in the securitized loan. However, raising one agent’s incentives and stake in the loan necessarily limits the other agent’s stake and incentives, leading to negative spillovers between the monitor’s and screener’s incentives. By contrast, when screening and monitoring are bundled and undertaken by the same agent, there are positive spillovers between screening and monitoring incentives. As a result, we find that it is optimal to bundle screening and monitoring in order to maximize positive incentive spillovers and synergies between these two tasks and reduce credit risk. According to our model, bundling is particularly beneficial for high quality borrowers—again providing a rationale for banks’ focus on this segment of credit markets—and when the costs of screening and monitoring are low. We also find that effort dynamics are different under separating and bundling. Notably, monitoring increases over time under separate tasks, while it decreases over time under bundling, with important effects on credit risk.

Our paper relates to the large banking literature on screening and monitoring. Most models in this literature are static; see e.g. Diamond (1984), Gorton and Pennacchi (1995), Holmstrom (1989), or Parlour and Plantin (2008). As a result, they do not explicitly distinguish between monitoring after loan origination and screening of loans at origination and cannot investigate the
dynamics of incentives and loan sales and their effects on credit risk. Our paper advances this literature by filling this important gap. Following early contributions by Sufi (2007) and Ivashina (2009), a growing empirical literature examines the effects of the loan stake of the lead arranger in syndicated loans on screening and monitoring (see e.g. Benmelech et al. (2012), Wang and Xia (2014), Bord and Santos (2015), or Gustafson et al. (2021)). Most of these studies proxy skin in the game by the originator’s initial stake in the loan. This literature has recently focused on loan sales after origination and their effects on incentives and credit risk (see e.g. Lee, Liu, and Stebunovs (2022) or Blickle et al. (2022)). Our model rationalizes such sales as part of an optimal contract between originators and outside investors. In particular, we demonstrate that, as screening occurs only once at origination while monitoring occurs after origination, moral hazard over screening and monitoring have different implications for incentive provision and credit risk.

From a modeling perspective, our paper builds on the literature that studies dynamic contracts in continuous time, starting with DeMarzo and Sannikov (2006) and Biais, Mariotti, Plantin, and Rochet (2007). In this literature, Piskorski and Westerfield (2016), Malenko (2019), and Gryglewicz and Mayer (2021) analyze incentive provision with optimal dynamic contracts and monitoring. Halac and Prat (2016) and Varas, Marinovic, and Skrzypacz (2020) characterize optimal monitoring in dynamic settings but do not focus on optimal contracts. None of these papers studies moral hazard over both screening and monitoring in credit markets. In a related paper, Hartman-Glaser, Piskorski, and Tchistyi (2012) study optimal securitization and screening of mortgages under moral hazard. In their model, the optimal contract features a single payout to the agent when sufficient time has elapsed after origination. While their optimal contract cannot be implemented using standard securities, it can be approximated by the first loss piece. Malamud, Rui, and Whinston (2013) and Hoffmann, Inderst, and Opp (2021) generalize the setting of Hartman-Glaser et al. (2012) by allowing for more general preferences and more general sources of uncertainty, respectively. Unlike ours, these papers do not model screening and monitoring and, as a result, cannot study optimal dynamic incentive provision in corporate loans. We show that the combination of screening and monitoring moral hazard implies a level of retention that gradually decreases over time. That is, with moral hazard over both screening and monitoring, the optimal contract is both about when the loan originator gets paid and what piece of the underlying loans it retains.

Hu and Varas (2021) study optimal intermediary financing in a setting in which the intermediary monitors the pool of loans she originates but is subject to moral hazard. The intermediary cannot commit to retaining a stake within the originated loans and thus sells off her retained stake over time, leading to a similar commitment problem as in, e.g., DeMarzo and He (2021). Our paper differs from theirs mainly in the following three aspects. First, we study an optimal contracting
problem without commitment frictions. While the optimal contract can be implemented via the agent’s retention of a stake within the pool of loans, the agent can perfectly commit to the optimal (dynamic) retention level. Second, Hu and Varas (2021) features no screening at loan origination. Third, we highlight the effects of debt maturity on the bank’s ability to commit. In particular, short maturity undermines the bank’s commitment, thereby providing a rationale for the use of long-term debt as a commitment device when the issuer is subject to agency conflicts.

Section 1 presents the model and discusses the contracting problem. Section 2 solves the model and derives the optimal contract. Section 3 discusses implications of our analysis. Sections 5 and 4 assess the effects of credit ratings and loan maturity on incentives and credit risk. Section 6 analyzes whether screening and monitoring should be bundled. Section 7 concludes. Technical developments are gathered in the Appendix.

1 Model Setup

Time $t$ is continuous and defined over $[0, \infty)$. A bank (the agent or “she”) originates a loan that can be sold to competitive outside investors (the principal or “they”). We start by considering that the loan has infinite maturity; Section 3 considers the model with finite maturity loans. The loan promises a constant flow payoff (coupon payments) normalized to 1 up to its default, which occurs at the random time $\tau$. The default time $\tau$ arrives according to a jump process $dN_t \in \{0, 1\}$ with (endogenous) intensity $\lambda_t > 0$ at time $t$, where $\tau := \inf\{t \geq 0 : dN_t = 1\}$. That is, over a short period of time $[t, t + dt)$, the loan defaults with probability $E dN_t = \lambda_t dt$.

The default rate $\lambda_t$ depends on the agent’s screening effort $q$ at time $t = 0$ and the agent’s monitoring effort $a_t$ at time $t \geq 0$. Specifically, the default intensity at time $t$ is given by

$$\lambda_t = \Lambda - a_t - q$$

(1)

where $\Lambda > 0$ captures the intrinsic risk (default intensity) of the loan. Screening and monitoring efforts are bounded, in that $q \in [0, \bar{q}]$ and $a_t \in [0, \bar{a}]$ with $\Lambda > \bar{a} + \bar{q}$. The bounds $\bar{a}$ and $\bar{q}$ are necessary to ensure that the instantaneous default probability $\lambda_t$ is well-defined and positive. Unless otherwise mentioned, we focus on parameter configurations that lead to optimal efforts $a_t \in [0, \bar{a})$ and $q \in (0, \bar{q})$, so the upper bound does not bind. The expected time to default at time $t$ is given by

$$\bar{\tau}_t = \int_t^\infty e^{-\int_t^s \lambda_u du} ds,$$

(2)

which reflects credit quality. In particular, a high value of $\bar{\tau} := \bar{\tau}_0$ at time $t = 0$ corresponds to low
credit risk, while a low value of $\bar{\tau} = \tau_0$ corresponds to high credit risk.

Screening entails a cost $\frac{1}{2}\kappa q^2$ at time zero. Monitoring entails a flow cost $\frac{1}{2}\phi a^2_t$ at time $t \geq 0$. Screening and monitoring efforts are unobservable to the principal and not contractible, giving rise to moral hazard. We do not impose any restrictions on the relation between screening and monitoring. Notably, we do not make any assumptions on whether screening and monitoring efforts are substitutes or complements. According to equation (1) screening and monitoring affect the instantaneous default rate $\lambda_t$ in a symmetric and independent way. If the bank decides to shirk on either task, the loan will have a higher default rate. Also notice that while they both reduce default risk, monitoring and screening differ in two ways. First, screening occurs once at time $t = 0$, whereas monitoring occurs frequently, specifically, at any point in time $t \geq 0$ up to default. Second, the effect of screening is more persistent than that of monitoring, where we consider for tractability that monitoring $a_t$ has purely transitory impact.

Both the principal (investors in the syndicate) and the agent (the bank) are risk neutral. The principal discounts cash flows at rate $r > 0$. The agent is more impatient and discounts cash flows at rate $\gamma > r$. The difference in discount rates may reflect the bank’s credit constraints or regulatory capital requirements, as in DeMarzo and Duffie (1999), or differences in financial constraints or risk-aversion, as in DeMarzo and Sannikov (2006).

Due to the discount rate differential $\gamma - r > 0$, there are gains from selling the loan—or a security whose payoff depends on loan performance—to outside investors, a process that works as follows. At inception, the bank designs a financial contract or, equivalently, a security $C$ that is sold to competitive investors at price $P_0$. The contract $C = \{dC_t, \hat{a}_t, \hat{q}\}$ represents a claim on the loan originated by the bank and stipulates a profit-sharing rule $dC_t$ of the overall loan payments $1dt$, so that the bank receives $dC_t$ and the investors receive $1dt - dC_t$ dollars over each time interval $[t, t + dt]$. The contract $C$ also stipulates monitoring efforts $\hat{a}_t$ (for all $t \geq 0$) and screening effort $\hat{q}$. We focus on incentive compatible contracts that induce actual monitoring (screening) effort $a_t$ ($q$) to coincide with contracted monitoring effort $\hat{a}_t$ ($\hat{q}$) and screening efforts, that is, $\hat{a}_t = a_t$ and $\hat{q} = q$. Unless necessary, we do not explicitly distinguish between contracted and actual effort levels.

Both the principal and the agent are protected by limited liability. That is, the principal’s (the agent’s) continuation payoff from following the contract $C$ must at any time exceed the principal’s (the agent’s) outside option, which we normalize to zero. Finally, while we do not impose any explicit constraints on the transfers $dC_t$, we show later that optimal transfers satisfy $dC_t \in [0, 1dt]$, so the bank receives positive payouts $dC_t \geq 0$ over each time interval $[t, t + dt]$. Conversely, under the optimal security (contract) $C$, investors receive $1dt - dC_t \geq 0$. 

8
Contracting problem

In what follows, \( t = 0^- \) denotes the time just before screening effort is chosen, and \( t = 0 \) is the time just after screening effort is chosen. At time \( t = 0^- \), the principal and the agent sign a contract \( C \), after which the agent chooses her screening effort \( q \). Notably, given the contract \( C \), the agent chooses screening effort \( q \) and monitoring effort \( \{a_t\} \) to maximize the expected present value of private profits

\[
W_0^- = \max_{q,\{a_t\}} E \left[ \int_0^\infty e^{-\gamma t} \left( dC_t - \frac{\phi a_t^2}{2} dt \right) \right] - \frac{kq^2}{2},
\]

where the subscript \( 0^- \) denotes values before screening effort is chosen. When buying the security from the bank (loan originator), outside investors have rational expectations regarding the bank’s incentives to exert screening and monitoring efforts.

It is natural to conjecture that the bank should not be rewarded for default in the optimal contract because this outcome indicates either poor monitoring, poor screening, or both. Hence, no positive payments should be made to the bank after time \( \tau \); that is, we should have \( dC_t \leq 0 \) for \( t \geq \tau \). In addition, limited liability rules out penalties for default, i.e., negative payments \( dC_t < 0 \) for \( t \geq \tau \). Altogether, we thus have that the payments to the bank satisfy \( dC_t = c_t dt \) for a flow compensation stream \( c_t \) at time \( t \geq 0 \).

The price that outside investors pay for a contract \( C \) at time \( t = 0^- \) is given by \( P_0^- = P_0 \) where the time-\( t \) price of the security is

\[
P_t = E_t \left[ \int_t^\tau e^{-r(s-t)} (1 - c_s) ds \right] = \int_t^\infty e^{-r(s-t)} \int_t^\tau \lambda_u du (1 - c_s) ds.
\]

In equation (4), the second equality integrates the default intensity \( \lambda_s \) over the relevant time interval. The bank receives \( P_0 \) dollars at time \( t = 0^- \) from selling the security to investors, in that \( dC_0^- = P_0 \). As outside investors are competitive, the bank can extract all the surplus and therefore chooses the security that maximizes total initial surplus \( F_0^- := W_0^- + P_0 \) at time \( t = 0^- \). That is, the bank solves

\[
\max_C F_0^-,
\]

taking into account her own moral hazard problem and the limited liability constraints.

Under the contract \( C \), the agent’s continuation payoff at time \( t \geq 0 \) is

\[
W_t := E \left[ \int_t^\tau e^{-\gamma(s-t)} \left( c_s - \frac{\phi a_s^2}{2} \right) ds \right] = \int_t^\infty e^{-\gamma(s-t)} \int_t^\tau \lambda_u du \left( c_s - \frac{\phi a_s^2}{2} \right) ds,
\]
where the second equality integrates the default intensity $\lambda_s$ over the relevant time interval. In this equation, $W_t$ is the expected, discounted value of the bank’s future payouts, adjusted for the cost of effort. As such, $W_t$ captures the value of the bank’s deferred compensation (deferred payouts). Because $P_t$ in (4) and $W_t$ in (6) can be expressed as deterministic integrals after integrating out the random default event and because the optimal contract dynamically maximizes total surplus $F_t = W_t + P_t$, the dynamic optimization problem (5) can be formulated as a deterministic problem. Unless otherwise mentioned, we adopt the deterministic formulation of problem (5).

2 Model solution

2.1 Incentives for screening and monitoring

We now turn to characterizing the bank’s incentives for screening and monitoring and, hence, the optimal effort levels $q$ and $\{a_t\}$. To begin with, let us fix screening effort at $q$ and analyze monitoring incentives given $q$. Limited liability requires that $W_t \geq 0$ for all $t \geq 0$, as otherwise, the bank would be better off leaving the contractual relationship. Owing to limited liability, outside investors do not receive payments from the agent in default. As a consequence, the agent only loses her claim on future payments, i.e., her continuation payoff $W_t$, at the time of default. With her monitoring activity, the agent controls the probability of default or equivalently the probability of losing future payments $W_t$ over the next instant, which is given by $\lambda_t dt = (\Lambda - a_t - q_t) dt$. Thus, the agent’s optimal monitoring effort is

$$a_t = \arg \max_{a_t \in [0, \bar{a}]} \left\{ -(\Lambda - a_t - q)W_t - \frac{\phi a_t^2}{2} \right\} = \arg \max_{a_t \in [0, \bar{a}]} \left\{ a_t W_t - \frac{\phi a_t^2}{2} \right\}.$$ 

As we focus on monitoring effort satisfying $a_t \in [0, \bar{a})$ and $W_t \geq 0$ (limited liability), the bank’s optimal monitoring effort is

$$a_t = \frac{W_t}{\phi}. \quad (7)$$

Equation (7) describes the incentive constraint for monitoring effort, in that incentive compatibility requires $a_t = a_t = \frac{W_t}{\phi}$ for all $t \geq 0$. According to equation (7), higher deferred payments $W_t$ increase the agent’s exposure to default risk and induce higher monitoring effort $a_t$. Therefore, deferred payments offer a trade-off. On the one hand, they provide monitoring incentives. On the other hand, they are costly due to the agent’s relative impatience ($\gamma > r$).

While monitoring $a_t$ impacts the instantaneous default intensity $\lambda_t$ at a single point in time $t$, screening $q$ affects all future default intensities $\{\lambda_t\}_{t \geq 0}$ and therefore the entire sequence of expected
payments, encapsulated in $W_0 = W_0(q)$. Note that we now explicitly recognize the dependence of $W_0$ on screening effort $q$ that is chosen “just before” time $t = 0$ at time $t = 0^-$. The agent chooses screening effort to maximize $W_0^-$ which is the value of her claim after screening is chosen, $W_0(q)$, net of the screening effort cost, $\frac{kq^2}{2}$:

$$\max_q \left( W_0(q) - \frac{kq^2}{2} \right). \quad (8)$$

Solving (8) for the optimal screening effort yields the incentive condition for screening effort $q$

$$\frac{\partial}{\partial q} W_0(q) = kq. \quad (9)$$

Lemma 1 below derives a condition such that the first-order condition (9) is sufficient for incentive compatibility, in that the first-order approach is valid.

**Lemma 1.** Suppose that the model parameters satisfy

$$\kappa \geq \frac{2}{(\gamma + \Lambda - \bar{a} - \bar{q})^2(r + \Lambda - \bar{a} - \bar{q})}. \quad (10)$$

Incentive conditions (7) and (9) hold and uniquely pin down the agent’s monitoring and screening effort. The incentive conditions (7) and (9) are sufficient and the first-order approach is valid.

Throughout the paper, we assume that condition (10) in Lemma 1 is met. In addition, we assume that:

$$\kappa \geq \frac{\phi \bar{a}}{q(\gamma + \Lambda - \bar{a} - \bar{q})}, \quad (11)$$

which is needed in the verification proof of the optimal contract in the proof of Proposition 2.

Let $V_t$ denote the agent’s gain from a marginal increase in $q$ measured from time $t$ onward, i.e.,

$$V_t = \frac{\partial}{\partial q} W_t(q).$$

The incentive condition in equation (9) can then be written as

$$q = \frac{V_0}{\kappa}. \quad (12)$$

That is, $V_t$ captures the agent’s screening incentives at time $t$ and, because screening effort is chosen at time $t = 0^-$, the value of $V_0$ determines the amount of screening $q$ exerted by the agent. Notably, equation (12) describes the incentive condition for screening effort, in that incentive compatibility for screening requires $q = \bar{q} = \frac{V_0}{\kappa}$.
Next, we characterize the dynamics of the agent’s monitoring and screening incentives $W_t$ and $V_t$. We can differentiate (6) with respect to time and obtain\(^5\)

$$\dot{W}_t := \frac{dW_t}{dt} = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - c_t. \quad (13)$$

To derive the law of motion of $V_t$, first calculate $\frac{dV_t}{dt} = \frac{d}{dt} \frac{\partial W_t}{\partial q} = \frac{\partial}{\partial q} \frac{dW_t}{dt}$. Next, note that at any point in time $t$, the bank’s continuation payoff can be written as the dynamic optimization problem:

$$\gamma W_t = \max_{a_t \in [0, \bar{a}]} \left( c_t - \frac{\phi a_t^2}{2} - \lambda_t W_t + \dot{W}_t \right) \quad (14)$$

which follows after rearranging (13) and accounting for the optimization over $a_t$. Using the envelope theorem, we can differentiate both sides of (14), evaluated under the optimal control $a_t$, with respect to $q$ to obtain\(^6\)

$$\dot{V}_t := \frac{dV_t}{dt} = (\gamma + \lambda_t)V_t - W_t. \quad (15)$$

Note that because screening effort $q$ is neither observable nor contractible, an unobserved change in screening effort $q$ cannot affect contracted flow payments at time $t$, in that $\frac{\partial c_t}{\partial q} = 0$ in the above calculation. Integrating the ordinary differential equation (15) over time $t$ yields the following expression

$$V_t = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} W_s ds. \quad (16)$$

Expression (16) shows that screening incentives $V_t$ decrease with the loan default rate. As a result, they are stronger if the loan is expected to default later, in which case screening effort has a longer lasting impact. Importantly, monitoring incentives by means of deferred payouts $W_t$ pin down the evolution of screening incentives $V_t$. That is, screening and monitoring incentives are closely linked and interact with each other. Higher $W_t$ exposes the agent’s compensation more strongly to loan performance and therefore motivates screening. In addition, higher $W_t$ boosts monitoring $a_t$, which delays default and strengthens screening incentives.

### 2.2 Optimal contract

In this section, we solve the model and characterize the optimal contract between the loan originator (the bank) and outside investors.

---

\(^5\)For a general payout process $dC_t$, it follows similarly that $dW_t = (\gamma + \lambda_t)W_t dt + \frac{\phi a_t^2}{2} dt - dC_t$. Under smooth payouts $dC_t = c_t dt$, this law of motion then simplifies to (13).

\(^6\)The first order condition with respect to monitoring reads $\frac{\partial W_t}{\partial a_t} = 0$, so that by the envelope theorem, $\frac{\partial}{\partial q} \frac{\partial W_t}{\partial a_t} = 0$. 

---
2.2.1 Benchmark: observable and contractible screening

To highlight the differences between monitoring and screening incentives more thoroughly, we start by studying the "second best" benchmark in which screening is not subject to moral hazard, in that \( q \) is publicly observable and contractible.

To solve the model under this benchmark, we first fix the screening level \( q \). We conjecture (and verify) that the optimal contract is stationary and features constant flow payments to the manager \( c_t = c > 0 \) until default, so that \( \dot{W}_t = 0 \) and \( W_t = W = W^B(q) \) for all \( t \). Inserting \( \dot{W}_t = 0 \) into equation (13) yields

\[
c = (\gamma + \Lambda - a - q)W + \frac{\phi a^2}{2}.
\]  

Equation (17) implies a one-to-one mapping between \( c \) and \( W \). As a result, controlling \( c \) is equivalent to controlling \( W \) and we can treat \( W \) as a choice variable instead of \( c \). Given screening effort \( q \) and constant monitoring effort \( a \), the default rate is constant and equal to \( \Lambda - a - q \), and the price of the security becomes:

\[
P^B(q) = \frac{1 - c}{r + \Lambda - a - q}.
\]  

Expression (18) is the discounted stream of flow payouts to outside investors, \( 1 - c \), where the (constant) default rate \( \Lambda - a - q \) augments the discount rate \( r \).

Next, note that given a screening level \( q \), the optimal monitoring effort \( a \) (and equivalently optimal deferred compensation \( W = \phi a \)) is chosen to maximize total surplus after screening is chosen, \( F^B(q) = P^B(q) + W \). Using equations (17) and (18), we get that the bank solves

\[
F^B(q) = \max_{W \in [0,F^B(q)]} \left( \frac{1}{r + \Lambda - a - q} - \frac{(\gamma - r)W}{r + \Lambda - a - q} - \frac{\phi a^2}{2} \right),
\]

where the choice of \( W \) determines monitoring effort \( a \) via equation (7), in that \( a = W/\phi \). Limited liability requires that both the agent’s continuation payoff \( W \) and the principal’s continuation payoff \( F^B(q) - W \) exceed zero, leading to \( W \in [0,F^B(q)] \). Equation (19) shows that the surplus \( F^B(q) \) consists of the value of the loan repayments minus agency and direct cost of monitoring and screening. Because the bank is subject to moral hazard, it must retain a stake \( W \), which generates agency cost due to its relative impatience, \( \gamma > r \). The maximization problem in (19) yields optimal levels of monitoring effort and deferred compensation, \( a^B(q) \) and \( W^B(q) \), given a fixed level of screening \( q \), whereby \( W^B(q) < F^B(q) \) and the principal’s limited liability constraint never binds.
Using (16), we can also calculate

\[ V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}. \]  

Equation (20) characterizes the agent’s screening incentives under the second best solution and plays an important role in the solution with non-contractible screening.

Finally, we can optimize \( F^B(q) \) over \( q \) to determine the optimal screening level in this second best benchmark:

\[ q^B = \arg \max_{q \in [0,\bar{q}]} \left( F^B(q) - \frac{\kappa q^2}{2} \right) \]  

determining second best screening effort \( q^B \), second best monitoring effort \( a^B(q^B) \), and second-best deferred payouts \( W^B(q^B) \). The optimal screening effort \( q^B \) solving (21) is then the solution to the first order condition

\[ \kappa q^B = \frac{F^B(q^B)}{r + \Lambda - a^B(q^B) - q^B}. \]  

We summarize our findings in the following proposition.

**Proposition 1** (Moral hazard over monitoring). Suppose that screening effort \( q \) is contractible, so that there is no moral hazard with respect to screening. At the optimum, the following holds:

1. For any choice of \( q \), monitoring effort \( a^B(q) \), payouts \( c^B(q) \), and deferred payouts \( W^B(q) \) are constant over time and are jointly characterized via (7), (17), and (19). The continuation payoff satisfies \( W^B(q) < F^B(q) \). Optimal monitoring effort \( a^B(q) \) increases with \( q \).

2. The optimal choice of screening effort, denoted by \( q^B \), is determined in (21) and solves the first order condition (22).

### 2.2.2 Moral Hazard over Screening and Monitoring

We now assume that \( q \) is unobservable to investors and consider the full contracting problem with moral hazard over both screening and monitoring. We solve this problem in two steps. As before, we first fix screening effort \( q \) and solve the continuation problem for \( t \geq 0 \). We then determine the optimal level of screening \( q = q^* \), taking into account the solution to the continuation problem.

Given levels of monitoring \( a \) and screening \( q \), we can rewrite the total surplus at time \( t \) (which
is the time-$t$ value of the bank’s objective in (5) as:\footnote{For a derivation, take $F_t = P_t + W_t$ in the first line of (23) and take the derivative with respect to time, $t$, to get

$$\dot{F}_t = \left(r + \lambda_t\right)P_t - 1 + c_t + (\gamma + \lambda_t)W_t - c_t + \frac{\phi a^2}{2} = \left(r + \lambda_t\right)(P_t + W_t) - 1 + \frac{\phi a^2}{2} - (\gamma - r)W_t.$$}

$$F_t = \int_t^\infty e^{-r(s-t)} \int_t^s \lambda u du (1 - c_s) ds + \int_t^\infty e^{-\gamma(s-t)} \int_t^s \lambda u du \left(c_s - \frac{\phi a^2}{2}\right) ds$$

$$= \int_t^\infty e^{-r(s-t)} \int_t^s \lambda u du \left(1 - \frac{\phi a^2}{2} - (\gamma - r)W_s\right) ds.$$ \hspace{1cm} (23)

As $V_t$ and $W_t$ characterize the agent’s incentives and there is no other (relevant) source of uncertainty than the arrival of the loan default time $\tau$, the variables $V_t$ and $W_t$ summarize all payoff-relevant information. Thus, we can express the total surplus as a function of $V_t$ and $W_t$, in that

$$F_t = F(V_t, W_t).$$

In what follows, we omit time-subscripts, unless necessary.

The integral expression (23) implies that the total surplus $F(V, W)$ solves:\footnote{For a derivation conjecture that $F_t = F(V_t, W_t)$ in the first line of (23) and take the derivative with respect to time, $t$, to get

$$\dot{F}_t = (r + \lambda_t)P_t - 1 + c_t + (\gamma + \lambda_t)W_t - c_t + \frac{\phi a^2}{2} = (r + \lambda_t)(P_t + W_t) - 1 + \frac{\phi a^2}{2} - (\gamma - r)W_t.$$}

$$rF(V, W) = \max_{a,c} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V, W) \right\}$$

$$+ F_V(V, W)((\gamma + \lambda)V - W) + F_W(V, W)\left((\gamma + \lambda)W + \frac{\phi a^2}{2} - c\right),$$ \hspace{1cm} (24)

where $F_V(V, W) = \frac{\partial F(V, W)}{\partial V}$ and $F_W(V, W) = \frac{\partial F(V, W)}{\partial W}$. Equation (24) is solved subject to the incentive condition (7), the limited liability constraints, and the conjecture that payouts to the bank are smooth, in that $dC = c dt$. Note that it is always possible to stipulate that the bank receives a payout of $\Delta$ dollars, which leaves $V$ unchanged but changes $W$ by $-\Delta$ dollars.\footnote{If payouts to the bank are not smooth, then it follows similar to (13) that

$$dW_t = (\gamma + \lambda_t)W_t dt + \frac{\phi a^2}{2} dt - dC_t,$$

so a payout of $dC = \Delta$ dollars reduces $W$ by $\Delta$, that is, $dW = -\Delta$.}

That is, controlling payouts to the bank is equivalent to controlling $W$. As a result, we can formulate the dynamic optimization problem of the bank such that $W$ instead of $c$ enters the HJB equation (24) as a control variable. Optimal payouts to the bank are then defined as the residual that implements the optimal $W$; see Section 3.2.
The optimality of payouts $c$ requires that

$$\frac{\partial F(V, W)}{\partial c} = -F_W(V, W) = 0.$$  

Substituting $F_W(V, W) = 0$ back into (24), we can rewrite (24) as

$$rF(V) = \max_{a, W} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)((\gamma + \lambda)V - W) \right\},$$  

where (with slight abuse of notation) $F$ is a function of $V$ only and $W$ is a control. Equation (25) is solved subject to the incentive condition for monitoring effort (7), i.e., $W = \phi a$, and the principal’s and the agent’s limited liability conditions, i.e., $W \in [0, F(V)]$.

Moral hazard over screening and the provision of screening incentives distort the optimal choice of monitoring incentives away from the benchmark with contractible (observable) screening. However, because the optimal contract must provide appropriate screening incentives only at inception at time $t = 0^-$ and the provision of these incentives as well as the distortion of monitoring incentives are costly due to $\gamma > r$, these distortions decrease over time. That is, optimal monitoring $a_t$ and the total surplus $F_t$ derived under the optimal contract from time $t$ onward approach the respective levels of the benchmark with observable screening as $t$ tends to $\infty$, in that

$$\lim_{t \to \infty} (a_t, W_t, V_t, F_t) = (a^B(q), W^B(q), V^B(q), F^B(q)).$$

That is, as time $t$ tends to infinity, the state variable $V$ approaches $V^B(q)$ which is defined in (20). Expressed in terms of the state variable $V$, equation (25) is solved subject to the boundary condition

$$\lim_{V \to V^B(q)} F(V) = F^B(q).$$

Recall that $F^B(q)$ and $V^B(q)$ depend on screening effort $q$. In addition, screening effort $q$ is linked with $V_0$ via the screening incentive compatibility condition (12), leading to $V_0 = \kappa q$.

We also show in the Appendix that $\kappa q = V_0 > V^B(q)$ in optimum. Over time, $V$ drifts down to $V^B(q)$, in that $\dot{V}_t < 0$ with $\lim_{t \to \infty} \dot{V}_t = 0$. Thus, the state space can be characterized by the interval $(V^B(q), V_0]$. The value function is downward sloping, with $F'(V) < 0$ for $V \in (V^B(q), V_0]$. In addition, we are able to show that the value function is strictly concave.

Having characterized the model solution for $t \geq 0$ and given screening effort $q$, we are now in a position to solve optimal screening effort. By the incentive compatibility condition (12), the initial
value of screening incentives, \( V_0 \), determines optimal screening effort \( q = q^* \) so that

\[
q^* = \arg \max_{q \in [0, \bar{q}]} \left( F(V_0) - \frac{\kappa q^2}{2} \right) \quad \text{s.t.} \quad V_0 = \kappa q. \tag{27}
\]

The following proposition summarizes the properties of the optimal contract.

**Proposition 2 (Moral hazard over screening and monitoring).** In optimum, the following holds:

1. For any given \( q \), total surplus at time \( t \) is a function of \( V \) only, in that \( F_t = F(V_t) \). The value function \( F(V) \) solves (25) subject to condition (26). The value \( V^B = V^B(q) \) is given by \( V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q} \).

2. Optimal monitoring is characterized by the maximization in (25) subject to (7), with its solution (28).

3. Optimal screening effort \( q = q^* \) is characterized in (27).

4. In optimum, it holds that \( \kappa q = V_0 > V^B(q) \), and \( V \) drifts down (i.e., \( \dot{V}_t < 0 \)) to \( V^B(q) \), but never reaches \( V^B(q) \) (i.e., \( V_t > V^B(q) \)).

5. The value function \( F(V) \) strictly decreases in \( V \) on \( [V^B(q), V_0] \) with \( \lim_{V \to V^B(q)} F'(V) \leq 0 \), so that \( F'(V) < 0 \) for \( V > V^B(q) \). The value function is strictly concave (with \( F''(V) < 0 \) at all \( V > 0 \) at which \( F'(V) \) is differentiable).

6. Payouts to the agent are smooth and positive.

Figure 1 provides a numerical example of the optimal contract. For the numerical analysis, we normalize \( r = 0 \), and \( \Lambda = 1 \). This choice of parameters implies that, without monitoring and screening, the expected time to default is \( 1/\Lambda = 1 \) year and the loan has pre-effort (or intrinsic) value \( 1/(\Lambda + r) = 1 \). In addition, we set \( \gamma = 0.1 \) and \( \phi = \kappa = 9 \) to generate the desired trade-offs. Last, we pick \( \tilde{a} = 0.125 \) and \( \tilde{q} = 0.2 \) to satisfy conditions (10) and (11). Our parameter choice implies that screening and monitoring effort are interior at all times (i.e., the constraints \( a_t \leq \tilde{a} \) and \( q \leq \tilde{q} \) never bind). The model’s qualitative outcomes are robust to the choice of these parameters.

The three upper panels of Figure 1 plot total surplus \( F(V) \), monitoring \( a(V) \), and the agent’s flow payouts \( c(V) \) as functions of the state variable \( V \). Observe that flow payouts to the agent are always positive. Likewise, as \( c_t < 1 \) at any time \( t \geq 0 \), flow payouts to the principal are positive too. The lower three panels depict the agent’s screening incentives \( V_t \), total surplus \( F_t \), and monitoring effort \( a_t \) as functions of time \( t \) (for \( t < \tau \)). Observe that \( V_t, F_t \) and \( a_t \) decrease over time with a decreasing speed. Even though not displayed, flow payments to the agent \( c_t \) decrease over time as
Figure 1: **Optimal contract.** In the upper panels, the dashed red line denotes the $V_0$. In the lower panels, the dotted red line denotes the benchmark levels that are attained in the limit $t \to \infty$.

Flow payments $c(V)$ increases with $V$ and $V$ decreases over time. Importantly, the dynamics of the value function $F_t = F(V_t)$ and monitoring effort $a_t = a(V_t)$ are shaped by the optimal incentive provision for screening. As screening only occurs at time $t = 0$, screening incentives and therefore the agent’s exposure to loan performance are front-loaded, thereby inducing a monitoring effort that exceeds the benchmark level $a^B = a^B(q^*)$. Intuitively, the provision of screening incentives distorts monitoring incentives upward, which is costly and curbs total surplus. Over time, these distortions taper off, improving total surplus $F_t$ so that $\dot{F}_t > 0$. Due to $\dot{V}_t < 0$ and $\dot{F}_t = F'(V_t)\dot{V}_t < 0$, it follows that $F'(V_t) < 0$ and total surplus decreases with $V$ for $V \in (V^B(q), V_0]$.

3 **Incentive provision and implementation**

3.1 **Dynamics of incentives**

We start by analyzing optimal incentives. Optimal monitoring follows from the first-order condition of the Hamilton-Jacobi-Bellman equation (25):

$$a(V) = \frac{\text{Reduction of default risk}}{\text{Physical cost}} \cdot \frac{\text{Screening incentives (}>0)}{\text{Agency costs}} = \frac{\hat{F}'(V)}{\phi} - F'(V)(V + (i - (\gamma - r))\phi) \wedge \frac{F(V)}{\phi},$$

(28)
where \( a(V) = \frac{F(V)}{\phi} \) when the limited liability constraint \( F(V) = W(V) \) binds and \( \land \) denotes the “min” operator (i.e., \( x \land y = \min\{x, y\} \)). Optimal monitoring \( a(V) \) is determined by several factors. First, monitoring reduces default risk, but comes at physical costs. Second, monitoring incentives require deferring the agent’s payments, which implies that \( W > 0 \) and is costly due to the discount rate differential (i.e., \( \gamma > r \)), generating agency costs. Third, monitoring incentives are linked to ex-ante screening incentives \( V_0 \) via

\[
V_0 = \int_0^\infty e^{-\gamma t - \int_0^t \lambda s W_t ds} dt,
\]

in that stronger monitoring incentives at any time \( t > 0 \) increase screening incentives at time \( t = 0 \). This effect results from two separate forces: (i) more monitoring \( a_t \) reduces the default intensity \( \lambda_t \) and so increases the expected time to default; i(ii) more monitoring incentives require exposing the agent to loan performance by raising \( W_t \), which also improves screening incentives. This effect is positive and, all else equal, increases monitoring effort and incentives above the benchmark level \( a^B = a^B(q^*) \), as illustrated in Figure 1. As screening is only performed at time \( t = 0 \), its benefits for the agent, as captured by the agent’s screening incentives \( V_t \) in equation (16), decrease over time within the optimal contract, converging to the level \( V^B = V^B(q^*) \) (see Figure 1). Because the strength of screening and monitoring incentives are linked, the agent’s monitoring incentives and so her monitoring also decrease over time. In the limit \( t \to \infty \) (i.e., \( V_t \to V^B = V^B(q^*) \)), \( a(V) \) approaches \( a^B(q^*) \). As a consequence, the instantaneous default rate \( \lambda_t \) capturing credit risk increases over time. Formally, because the value function is strictly concave, monitoring effort \( a(V) \) decreases with \( V \) and decreases over time due to \( \dot{V} < 0 \).

The following corollary summarizes our findings:

**Corollary 1.** Suppose that \( W(V) < F(V) \). Then, monitoring effort \( a(V) \) and the agent’s deferred compensation \( W(V) = \phi a(V) \) increase with the marginal benefits of screening \( V \), in that \( a'(V) > 0 \). Because \( V \) decreases over time, monitoring effort and deferred compensation decrease over time, with \( \lim_{t \to \infty} a_t = a^B \).

We now study how monitoring and screening effort change with model parameters of interest, i.e., \( \Lambda \) (credit risk), \( \kappa \) (screening cost), \( \phi \) (monitoring cost), and \( \gamma \) (bank’s discount rate/cost of capital). Figure 2 plots optimal screening \( q^* \) as well as monitoring efforts against the cost parameters \( \phi \) and \( \kappa \) and the baseline default intensity \( \Lambda \). As monitoring effort changes over time, we plot monitoring effort \( a_t \) at three different times, i.e., \( t = 0, t = 5, \) and \( t \to \infty \), to better account for the dynamics of monitoring effort. The left and center panels of Figure 2 (i.e., Panels A, B, E, and F) illustrate that monitoring effort \( a_t \) and screening effort \( q \) decrease with both the physical costs
Figure 2: Comparative Statics. This figure plots monitoring effort $a_t$ at $t = 0$ (solid black line), at $t = 5$ (dotted red line), and $t \to \infty$ (dashed yellow line) and screening effort $q^*$ against the parameters $\phi, \kappa, \Lambda, \text{and } \gamma$.

of monitoring and screening, $\phi$ and $\kappa$. That is, screening and monitoring efforts are complements. The underlying mechanism is that screening and monitoring incentives are determined and linked by the agent’s deferred compensation. Thus, the provision of strong screening incentives implies and requires strong monitoring incentives, while strong monitoring incentives boost the agent’s screening incentives. As a result, when the cost of screening $\kappa$ increases, it becomes optimal to reduce contracted screening effort, leading to lower screening incentives and, as such, to lower monitoring (incentives). Likewise, when the cost of monitoring $\phi$ increases, it becomes optimal to curb contracted monitoring and monitoring incentives, leading to lower screening (incentives).

Panels C and G of Figure 2 also illustrate that a decrease in the quality of the borrower (or in the quality of the loan), as reflected by the higher baseline default intensity $\Lambda$, leads to a decrease in monitoring and screening, due to lower marginal benefits of monitoring and screening. That is, our paper suggests a two-way relation between credit risk and lenders’ screening and monitoring. Notably, a worsening of credit quality leads to lax monitoring and screening, which in turn exacerbates credit risk. Our model, therefore, provides a rationale for the segmentation observed in credit markets. According to our analysis, banks that exert high screening and high monitoring
(e.g., via loan covenants) typically finance high quality (low $\Lambda$) borrowers with high priority loans. By contrast, private equity and online lenders finance lower quality (high $\Lambda$) borrowers with low priority debt instruments. Our analysis also suggests that when screening is more lax, monitoring should also be more lax. It is therefore consistent with the trend observed in the leveraged loan market, in which the incidence of including covenants is decreasing and where more than 80% of outstanding loans in 2020 are covenant light according to S&P Global Market Intelligence.\footnote{A similar trend can be observed in the corporate bond market in which we observe both a declining quality of borrowers and a decrease in the usage of bond covenants. See e.g. Celik, Demirtaş, and Isaksson (2019). We show in section 5 that a similar result obtains in the presence of a credit rating.}

Finally, Panels D and H of Figure 2 show that, as the bank’s cost of capital (discount rate) $\gamma$ increases, it becomes more costly to delay payouts to the bank and to provide incentives, so that screening and monitoring efforts decrease with $\gamma$.

### 3.2 Implementation

This section shows that the optimal contract can be implemented by having the bank retain a time-decreasing share of the loan. At origination (i.e., at time $t = 0$), the bank (the agent) retains a fraction $\beta_0$ of the loan and sells a fraction $1 - \beta_0$ to competitive outside investors. After origination at times $t \geq 0$, the bank smoothly sells off its stake $\beta_t$ so that the fraction of the loan retained decreases over time. That is, the agent owns a fraction $\beta_t$ of the loan at time $t$, where $\beta_t$ is adjusted to provide appropriate incentives $W_t$.

A per-unit claim on the loan pays the loan rate 1 up to default at time $\tau$ and therefore has a competitive price

$$D_t = \int_{t}^{\infty} e^{-r(s-t)} \int_{t}^{s} \lambda_u du ds,$$

at any time $t \geq 0$. $D_t$ is linked to credit risk via the instantaneous default intensities $\{\lambda_s\}_{s \geq t}$.

Over a short period of time $[t, t+dt]$, the agent receives $\beta_t dt$ in interest payments from the loan. In addition, she sells the loan at rate $-\dot{\beta}_t dt$, which yields trading revenues $-\dot{\beta}_t D_t dt$. Therefore, matching the payoffs of the optimal contract requires that:

$$\beta_t - \dot{\beta}_t D_t = c_t.$$  

Note that as the HJB equation (25) determines optimal monitoring incentives, and hence optimal deferred compensation $W_t = W(V_t)$, the agent’s payouts are implicitly characterized in (13). That is, we can solve (13) to get

$$c_t = (\gamma + \lambda_t) W_t + \frac{\phi a_t^2}{2} - \dot{W}_t > 0.$$  

\footnote{A similar trend can be observed in the corporate bond market in which we observe both a declining quality of borrowers and a decrease in the usage of bond covenants. See e.g. Celik, Demirtaş, and Isaksson (2019). We show in section 5 that a similar result obtains in the presence of a credit rating.}
Figure 3: **Implementation of the optimal contract.** The dotted red line depicts the $V = V_0$.

This equation, together with equation (30), implies that

$$\beta_t - \dot{\beta}_t D_t = (\gamma + \lambda_t) W_t + \frac{\dot{\phi} a_t^2}{2} - \dot{W}_t,$$

(32)

which pins down the rate $\dot{\beta}_t$ at which the agent sells off her stake (see also Appendix D.2).

Figure 3 presents a numerical example of the implementation of the optimal contract and plots the (per-unit) value of the loan and the issuer’s stake against $V$ (upper two panels) and against time $t$ (lower two panels). As time passes, the agent sells her stake $\beta_t$ and monitoring incentives decrease, which increases default risk and decreases the (per unit) value of the loan $D_t$.

The following proposition summarizes our results:

**Proposition 3** (Implementation). *The optimal contract can be implemented as follows. The agent retains a fraction $\beta_t$ of the originated loan at time $t$, whereby a unit stake pays out a flow payoff of 1 dollars until liquidation at time $\tau$ and has a competitive time-$t$ price given by (29). Over time, the agent sells her stake according to (32).*

Finally, we examine how the two moral hazard problems over screening and monitoring affect contract design and implementation. For this purpose, it is instructive to discuss two benchmarks in more detail. First, consider that there is no moral hazard over screening in that $q$ is observable and contractible, but monitoring is subject to moral hazard. As shown in Section 2.2.1, the solution is time-stationary with constant monitoring $a^R(q) = W^B(q)/\phi$ up to liquidation, solving (7), and
optimal screening $q = q^B$, solving (21). Interestingly, the following corollary shows that the optimal contract can be implemented by requiring the agent to retain a constant share of the loan.

Corollary 2. Suppose that there is no moral hazard over screening, and $q$ is observable and contractible. Then, the optimal contract can be implemented by requiring the agent to retain a constant share $\beta^B \in [0, 1]$ of the loan.

Second, consider that screening is subject to moral hazard, but monitoring is not in that $a_t$ is contractible and observable. As before, total surplus can be expressed as a function of the agent’s screening incentives $V$ and solves the HJB equation (25). However, different from the baseline, the incentive constraint (7), linking $W$ and $a$, does not apply. As we show, the value function has slope $F'(V) \leq - (\gamma - r)$, so the maximization with respect to $W$ yields that $W = F(V)$. Thus, the agent receives the highest amount of incentives through deferred compensation as the principal’s limited liability constraint permits. Over time, $V$ drifts down and reaches zero at some finite time $\tau_0$. At time $\tau_0$, the agent receives a lumpy payout $dC = F(0)$. The optimal contract can then be implemented by requiring the agent to retain the entire loan until time $\tau_0$. At that time, the agent sells the entire loan competitive outside investors. This implementation maximizes the agent’s exposure to loan performance before time $\tau_0$ (while respecting the principal’s limited liability) and allows the agent to capitalize on total surplus $F_{\tau_0} = F(0)$ at time $\tau_0$.

While the setting without monitoring moral hazard resembles that of Hartman-Glaser et al. (2012), there is one important difference in that both the agent and the principal have limited liability. By adding a limited liability constraint on the principal’s side, we obtain that the optimal contract is implementable using standard securities, a result that does not obtain in Hartman-Glaser et al. (2012). The following proposition summarizes these findings.

Proposition 4 (Moral hazard over screening). Suppose that there is no moral hazard over monitoring in that $a_t$ is observable and contractible. In such environments:

1. The value function $F(V)$ solves the HJB equation (25) subject to the boundary condition $F'(0) = -(\gamma - r)$. For $V > 0$, we have $F'(V) < -(\gamma - r)$ and the value function is strictly concave, so that setting $W = F(V)$ is optimal.

2. Over time, $V = V_t$ drifts down and reaches 0 at time $\tau_0$. Optimal effort satisfies

$$a(V) = \frac{F(V) - F'(V) V}{\phi}$$

and increases with $V$. As a result, optimal monitoring decreases over time.
3. The optimal contract implies that the agent’s payouts satisfy $dC_t = 1dt$ for $t < \tilde{\tau}$, $dC_t = F(0)$ for $t = \tilde{\tau}$, and $dC_t = 0$ for $t > \tilde{\tau}$ where $\tilde{\tau} = \inf\{t > 0 : V_t = 0\}$. As such, the optimal contract can be implemented with the agent retaining the entire loan until time $\tilde{\tau}$. At time $\tilde{\tau}$, the agent sells the entire loan to competitive outside investors.

Corollary 2 and Proposition 4 have interesting implications for the relation between agency conflicts and the bank’s optimal level of skin in the game. Interestingly, the severity of moral hazard affects both the level and the dynamics of the agent’s retention. A surprising result of Proposition 4 is that less severe agency conflicts, i.e., removing moral hazard over monitoring, actually increase the bank’s optimal initial retention, as optimal initial retention in the baseline model with moral hazard over both tasks is smaller than one.

3.3 Optimal retention and retention dynamics

As the optimal contract between the loan originator and outside investors can be implemented by having the loan originator retain a time-decreasing stake in the loan, both (i) the (initial) retention level and (ii) the speed at which the bank sells its stake determine the strength of dynamic screening and monitoring incentives. We now study how loan characteristics (e.g., intrinsic credit risk $\Lambda$), as well as the costs of monitoring and screening $\phi$ or $\kappa$, affect initial retention and the selloff dynamics. To this end, Figure 4 plots the initial retention levels $\beta_0$ and a measure of the selloff speed, that is, $1 - \beta_T/\beta_0$, against $\kappa$, $\phi$, $\Lambda$, and $\gamma$. Notice that $1 - \beta_T/\beta_0$ is the fraction of its initial stake that the bank sells up to time $T$; thus, if $1 - \beta_T/\beta_0$ is high (low), the bank sells off its initially stake quickly (slowly). In Figure 4, we take $T = 3$. Other measures of selloff speed yield similar results.

Figure 4 reveals that as the cost of monitoring $\phi$, credit risk $\Lambda$, or the bank’s discount rate $\gamma$ increase, the initial retention level decreases and the selloff speed increases (see Panels B, C, F, and G), so that overall the bank’s incentives to screen and monitor decrease in line with Figure 2. The model, therefore, predicts that the bank initially retains a lower fraction of the originated loans when credit risk ($\Lambda$) is high—which could be the case when the loans are of bad quality or the loans’ underlying collateral is risky—or when moral hazard over monitoring ($\phi$) is severe, and sells these loans faster. This prediction is consistent with the evidence in Adelino, Gerardi, and Hartman-Glaser (2019), who find a strong relationship between mortgage performance and time to sale for privately securitized mortgages. Although the authors interpret their findings in the context of an adverse selection model, our results show that moral hazard generates similar patterns.\[\text{\textsuperscript{11}}\]

\[\text{\textsuperscript{11}}\text{In their model, a separating equilibrium emerges in which the time to sale of a mortgage increases in quality, a relationship often referred to as the skimming property. See also Daley and Green (2012).}\]
Importantly, our model has additional predictions about the initial retention level, which naturally lend themselves to empirical testing. The result that, quite surprisingly, the initial retention level $\beta_0$ is larger when moral hazard over monitoring is less severe is closely related to the finding of Proposition 4 that, absent moral hazard over monitoring, the bank’s initial retention level equals one. Besides, the initial retention level $\beta_0$ decreases with the cost of screening $\kappa$, but the selloff speed quantitatively does not change much with $\kappa$ (and is non-monotonic in $\kappa$).  

4 The Effects of Loan Maturity

In our baseline model, loans have infinite maturity. As screening and monitoring efforts have effects of different duration, loan maturity could have different effects on these two tasks. In this section, we follow Chen, Xu, and Yang (2021) and consider that the loan randomly matures with Poisson intensity $\delta > 0$. That is, ignoring default, the expected loan maturity is $1/\delta$. The baseline setting

\[\text{To get some intuition for this non-monotonic relationship, notice that when $\kappa$ is sufficiently low, then there is effectively no moral hazard over screening, so the optimal contract comes close to the one from the benchmark in Section 2.2.1 with only monitoring moral hazard and a constant level of retention (i.e., zero selloff speed), which mirrors the result from Corollary 2. Similarly, when $\kappa$ becomes sufficiently large, the contracted screening level tends to zero and the optimal contract only incentivizes monitoring, again featuring (approximately) constant retention.}\]
corresponds to the case $\delta = 0$, in which the loan has infinite maturity. Up to its maturity date, the loan makes coupon payments at rate 1. When the loan matures, the firm pays back the face value $F_t^\delta$. That is, at maturity, the game ends and the bank and outside investors exit and $F_t^\delta$ represents their joint terminal payoff.

With finite maturity, the total continuation surplus satisfies

$$F_t = \int_t^\infty e^{-(r+\delta)(s-t)-\int_t^s \lambda_u du} \left( 1 - \frac{\phi a^2}{2} - (\gamma - r) W_s + \delta F_s^\delta \right) ds. \quad (33)$$

This expression differs from that in the baseline model in (23) as the loan matures at rate $\delta$, leading to the terminal payoff $F_t^\ast$. At origination (i.e., at time $t = 0^-$), the bank solves

$$\max_c \left[ F_0 - \frac{\kappa q^2}{2} \right] \quad (34)$$

subject to the incentive constraints i) $V_0 = \kappa q$ (screening) and ii) $a_s = W_s/\phi$ (monitoring). The agent’s screening incentives at time $t = 0$ read

$$V_0 = \int_0^\infty e^{-(\gamma+\delta)t-\int_0^t \lambda_u du} W_t dt. \quad (35)$$

At the time of maturity, the bank exits and is no longer exposed to loan default risk, so its screening incentives fall to zero. This is also reflected in the law of motion of $V_t$ which becomes

$$\dot{V}_t = (\gamma + \delta + \lambda_t)V_t - W_t. \quad (36)$$

In contrast, loan maturity has no direct effect on monitoring incentives, as the impact of monitoring at time $t$ is instantaneous. As a result, the Hamilton-Jacobi-Bellman equation for total surplus before the loan matures becomes

$$(r + \delta) F(V) = \max_{a, W} \left\{ 1 + \delta F^\delta - \frac{\phi a^2}{2} - (\gamma - r) W - \lambda F(V) + F'(V)(\gamma + \delta + \lambda)V - W \right\}. \quad (37)$$

As in the baseline, the boundary condition

$$\lim_{V \to V_B(q)} F(V) = F^B(q) = \max_{W \in [0, F^B(q)]} \left( \frac{1 + \delta F^\delta}{r + \Lambda - a - q + \delta} - \frac{\gamma - r}{r + \Lambda - a - q + \delta} - \frac{\phi a^2}{r + \Lambda - a - q + \delta} \right)$$

applies. In addition, as in the baseline model, $V_0 > V_B(q)$, and optimal screening effort $q^\ast$ maxi-
minizes total initial surplus \( F_0^- = F(V_0) - \frac{\kappa q^2}{2} \) subject to the incentive constraint \( V_0 = \kappa q \) where

\[
V^B(q) = \frac{W^B(q)}{r + \delta + \lambda^B}.
\] (39)

To analyze the effects of finite loan maturity on outcome variables, we assume for simplicity that \( F^\delta_t = F_t \) (or \( F^\delta = F(V) \)), i.e., the random maturity event leaves the total loan value unchanged. At maturity, the bank is paid \( W_t \) and outside investors are paid \( F_t - W_t \) dollars. Therefore, there is no value effect associated with the random maturity event.\(^{13}\) This assumption reflects in reduced form the fact that the value of the loan is the same “just before” maturity and at maturity; in a model with a deterministic maturity date, this property would be called a value matching condition.\(^{14}\) To begin with, note that there are two differences compared with the HJB equation (25). First, the loan matures with intensity \( \delta \) and total surplus changes by \( F^\delta - F(V) \) when this happens (which is assumed to equal zero). Second, loan maturity affects screening incentives \( V \), so that \( \delta \) shows up in the flow term that multiplies \( F'(V) \). That is, according to (35), screening incentives \( V_0 \) decrease with \( \delta \) (i.e., increase with loan maturity \( 1/\delta \)). That is, short maturity weakens the bank’s commitment (to screening incentives). The lower level of commitment implies that total surplus at origination, \( F_0^- = F(V_0) - \frac{\kappa q^2}{2} \), increases with \( 1/\delta \) (see Figure 5). Our model, therefore, provides a rationale for the use of long-term debt in the presence of commitment and agency frictions at the bank (originator) level.

Figure 5 plots initial monitoring effort \( a_0 \) proxying average monitoring (Panel A), screening effort \( q^* \), and the expected time to default \( \tau \) (which is inversely related to credit risk) against average debt maturity \( 1/\delta \).\(^{15}\) Short maturity undermines the bank’s commitment to high powered screening incentives and therefore screening incentives as such. Moreover, the bank’s initial retention level \( \beta_0 \) increases as debt maturity decreases. Larger initial retention then leads to higher initial monitoring \( a_0 \) by the bank. Thus, screening increases and (initial) monitoring decreases with debt maturity \( 1/\delta \). Figure 5 also plots the expected time to default \( \bar{\tau} \) (which is inversely related to credit risk) as a function of debt maturity \( 1/\delta \) and shows that credit risk decreases as maturity increases (i.e.,

\(^{13}\)This assumption has no bearings on our key findings and is for mere simplicity; our results would remain qualitatively unchanged had we assumed different \( F^\delta_t \), for instance, \( F^\delta_t = K \) for a constant \( K \geq 0 \).

\(^{14}\)In reality, loans mature deterministically and this feature naturally holds: If the value of debt were different “just before” and at maturity, there would be arbitrage.

\(^{15}\)To facilitate a fair comparison of credit risk across different loan maturities, we also consider for finite loan maturity

\[
\tau := \int_0^\infty e^{-\beta_0^* \lambda u} du dt
\]

as an (inverse) measure of credit risk (see also (2)). That is, we calculate the expected time to default (at time \( t = 0 \)) conditional on the loans not maturing. This way, we eliminate the mechanical effect that short maturity naturally limits the timespan over which the loan is exposed to credit risk, which limits credit risk.
Figure 5: The effects of debt maturity on screening, monitoring, credit risk, and total surplus. The dotted red line depicts the outcomes with infinite debt maturity.

\[ \bar{\tau} \text{ increases with } 1/\delta \). That is, the effect of maturity on screening dominates that on monitoring. Total surplus also increases with debt maturity due to lower agency costs.

5 The effects of credit ratings and CLOs

One way for loan originators to reduce their share in the loans they originate is to use securitization, for example, by including CLOs in the syndicate. A feature that CLOs share is that each loan included in the deal gets rated. As we show next, credit ratings in our setting asymmetrically affect the moral hazard problems over screening and monitoring. This asymmetry can be so severe that introducing even perfectly informative and contractible credit ratings can increase credit risk.

To characterize the effects of credit ratings on credit risk and total surplus, we consider a setting in which the loan is rated once at origination, i.e., at time \( t = 0 \).\(^{16}\) To focus on the effects of credit ratings on incentives and credit risk, we assume that the rating agency perfectly observes the credit risk and reports it truthfully, in that the credit rating is publicly observable and contractible. In

\(^{16}\)This assumption captures the feature of the market that ratings are issued relatively infrequently. Assuming more frequent ratings would not eliminate the fundamental mechanism we identify in this section that credit ratings have different effects on the moral hazard problems related to screening and future monitoring.
our setting, the credit rating reveals the initial credit quality and screening effort \( q \) that is chosen at origination. That is, with a credit rating at time \( t = 0 \), screening effort becomes publicly observable and contractible (chosen at time \( t = 0 \)), which removes the moral hazard over screening at origination.\(^{17}\) Intuitively, the credit rating at origination generates screening incentives, as lax screening would lead to a low rating. Because the credit rating cannot condition on the actual levels of monitoring that are chosen after the rating, it does not directly affect the originator's monitoring incentives after the time of the rating. As a result, the benchmark model without moral hazard over screening described in section 2.2.1 can be seen as a model with credit ratings. Proposition 1 characterizes optimal screening and monitoring in this benchmark model.

Figure 6 illustrates the effects of credit ratings on outcome variables by plotting the percentage change in monitoring effort (first row), screening effort (second row), and initial retention (third row) at \( t = 0 \) due to a credit rating. As shown by the figure, the credit rating increases screening at origination but reduces monitoring \( a_0 \). The reason is that due to the screening incentives from credit ratings, the agent requires lower screening incentives through deferred payouts and therefore retains a lower share in the loan, leading to lower monitoring incentives. That is, while the credit rating increases the agent's incentives to screen loans at origination, it undermines her incentives to monitor them afterward. Intuitively, the credit rating at origination can be understood as a complement to the lender's screening, and as a substitute to her monitoring. Notably, Figure 6 (third row) shows that under all parameters considered, a credit rating reduces the bank's initial retention level. The intuition is that by removing the moral hazard problem over screening, the credit rating allows the bank to reduce its incentives-based exposure to the pool (and eliminate front-loading). In addition, and as shown in Proposition 1, the credit rating affects the optimal retention level and implies that the bank (loan originator) retains a constant stake in the loan.

Due to their opposite effects on screening and monitoring incentives, credit ratings may increase or decrease credit risk, depending on bank and borrower characteristics. Notably, Figure 6 shows that when the cost \( \kappa \) of screening is low and screening effort is \( q^* \) is high in the baseline model, a credit rating has small effect on the originator's screening, while reducing monitoring incentives. In this case, a credit rating leads to higher credit risk, as reflected in the shorter expected time to default \( \bar{\tau} \) (bottom row in Figure 6). As the cost of screening becomes larger, moral hazard over screening becomes more important and credit ratings lead to a decrease in default risk.

\(^{17}\)Recall that the principal and the agent sign a contract at time \( t = 0^- \), i.e., just before screening effort is chosen. The credit rating makes the choice of \( q \) publicly observable and contractible, so one can think of screening and credit rating occurring simultaneously. Another way to think about credit rating is as follows. The rating could also happen after screening effort is chosen: then, investors get their money back (and the contract is reneged) if the bank deviates from the promised screening effort, which makes screening effort contractible.
Figure 6: The effect of credit ratings on effort levels, retention, default risk, and total surplus. $\Delta y$ denotes the percentage change in the initial value of the outcome variable $y$ caused by a credit rating, where $y \in \{a_0, q^*, \beta_0, \tau\}$. Outcome variables are plotted as functions of the cost of monitoring $\kappa$, the cost of screening $\phi$, the raw default intensity $\Lambda$, and loan maturity $1/\delta$. 
6 Is it optimal to bundle monitoring and screening?

We have so far assumed that the loan originator is responsible for both screening and monitoring. In practice, screening and monitoring may be undertaken by separate entities. Some securitized loans are serviced by a third-party serving company and, depending on the specific arrangements, servicing can subsume monitoring activities. In these cases, the originator is in charge of screening and the servicer in charge of monitoring. An important question is therefore whether bundling or separating screening and monitoring affects incentives and credit risk. To address this question, we consider a setting in which monitoring and screening are conducted by two different agents (called the monitor and screener). To make the comparison with the baseline model sensible, we assume that the monitor and the screener have identical preferences. We denote the monitor’s continuation payoff by $W^m_t$ and the screener’s continuation payoff by $W^s_t$. Both the screener and the monitor are subject to moral hazard. In what follows, we provide the heuristic solution with the separation of the two tasks. Appendix D.5 provides the detailed solution.

As in the baseline model, monitoring effort is determined by the monitor’s incentive condition

$$a_t = \frac{W^m_t}{\phi}.$$  

Screening effort is determined by the screener’s incentive condition

$$q = \frac{V_0}{\kappa},$$

where $V_t = \frac{\partial}{\partial q} W^s_t$. Similarly, as in the model with finite maturity loans, we have that

$$V_0 = \int_0^\infty e^{-(\gamma + \delta)t - \int_0^t \lambda_t ds} W^s_t dt.$$  

In the benchmark without moral hazard over screening, there is no point providing screening incentives to the loan originator, so that $V^B(q) = 0$. As such, after the loan matures, the bank no longer receives screening incentives, and the continuation surplus becomes identical to $F^B(q)$ from (19). To incentivize screening at $t = 0$, it must be that $V_0 > 0$. $V$ then drifts down over time.

We can express total surplus as a function $F(V)$ that depends on the screener’s incentives $V$, while treating $W^m$ and $W^s$ as control variables for the dynamic optimization problem of the bank.
Similar steps as in the baseline model show that total surplus solves

\[(r + \delta)F(V) = \max_{a, W^m, W^s}\left\{ \mu - \frac{\phi a^2}{2} - (\gamma - r)(W^m + W^s) - \lambda F(V) + \delta F^\delta + F'(V)((\gamma + \lambda + \delta)V - W^s) \right\}, \tag{40} \]

where limited liability requires that \(W^m \in [0, F(V) - W^s]\) and \(W^s \in [0, F(V) - W^m]\) and incentive compatibility with respect to monitoring requires that \(W^m = a\phi\). In (40), \(F^\delta\) is the total surplus “just after” the loan matures (which occurs at rate \(\delta\)). The surplus function satisfies

\[\lim_{V \to 0} F'(V) = -\gamma - r\] and \(\lim_{V \to 0} F(V) = F_B(q)\). Besides, the value function is strictly concave, so that \(F'(V) < -\gamma - r\) for \(V > 0\). Owing to \(F'(V) < -\gamma - r\), the maximization in (40) with respect to \(W^s\) yields

\[W^s = F(V) - W^m.\]

Inserting this expression into (40) and simplifying leads to the ordinary differential equation

\[(\gamma + \delta)F(V) = \max_{a, W^m}\left\{ \mu - \frac{\phi a^2}{2} - \lambda F(V) + \delta F^\delta + F'(V)((\gamma + \lambda + \delta)V - F(V) + W^m) \right\}, \tag{41} \]

which is solved subject to \(a = W^m/\phi\). As in Section 4, the loan value remains unchanged when the loan matures, in that the bank and outside investors receive \(F^\delta = F(V)\) upon maturity.

As \(V\) approaches zero, the boundary condition \(\lim_{V \to 0} F'(V) = -(\gamma - r)\) applies. Due to (41) and \(a = W^m/\phi\), the condition \(\lim_{V \to 0} F(V) = -(\gamma - r)\) implies

\[\lim_{V \to 0} F(V) = F_B(q) = \max_{a \in [0, \bar{a}]} \left( \mu - \frac{\phi a^2}{2} - \phi a(\gamma - r) \right) \left( \frac{r + \lambda}{r} \right), \tag{42} \]

which is expression (19). The maximization in (41) with respect to monitoring effort yields

\[a(V) = \frac{\text{Reduction of default risk}}{\phi} \frac{\text{Screening incentives}(>0)}{-F'(V)V} \frac{\text{Screening disincentives}(<0)}{+F'(V)\phi} \land \frac{F(V)}{\phi}. \tag{43} \]

Note that when the limited liability constraint for the screener binds and \(W^s = 0\), then \(W^m = F(V)\) and therefore \(a(V) = \frac{F(V)}{\phi}\). As in the baseline model, optimal screening effort \(q^*\) maximizes total initial surplus \(F_0^- = F(V_0) - \frac{q^2}{2}\) subject to the incentive constraint \(V_0 = \kappa q\). We summarize our findings in the following proposition.
Proposition 5 (Unbundling screening and monitoring). Suppose that monitoring and screening are undertaken by two different agents, called the monitor and screener, who are otherwise identical. Denote the monitor’s continuation value by $W^m_t$, the screener’s continuation value by $W^m_t$, and the screener’s screening incentives by $V_t$. As in Section 4, consider “value matching” at maturity, i.e., $F^\delta = F(V)$. In such environments:

1. The value function $F(V)$ solves the HJB equation (41) subject to the boundary condition $F'(0) = - (\gamma - r)$. For $V > 0$, we have $F'(V) < -(\gamma - r)$ and the value function is strictly concave, so that setting $W^s = F(V) - W^m$ is optimal.

2. Over time, $V = V_t$ drifts down and reaches 0 at time $\tilde{\tau} < \infty$. In this case, the boundary condition $F'(0) = - (\gamma - r)$ applies. Optimal effort satisfies (43).

3. Optimal screening $q^*$ solves

$$\max_{q \in [0, \bar{q}]} \left( F(V_0) - \frac{\kappa q^2}{2} \right) \quad \text{s.t.} \quad V_0 = \kappa q.$$ 

Monitoring and screening are linked by positive and negative terms, denoted as screening incentives and disincentives in equation (43). On the one hand, monitoring reduces the likelihood of default, leading to a longer lasting impact of screening and therefore to stronger screening incentives. On the other hand, stronger monitoring incentives require raising the monitor’s deferred compensation, which, in turn, requires lowering the screener’s deferred compensation to satisfy the limited liability constraints. This second effect leads to negative spillovers between monitoring and screening incentives. In contrast, when one agent is responsible for both monitoring and screening, monitoring effort satisfies (28) and monitoring unambiguously boosts screening incentives, leading to positive spillovers between monitoring and screening incentives. As a result, while bundling monitoring and screening leads to positive synergies, separating these two tasks can lead to negative synergies. Accordingly, bundling screening and monitoring leads to higher screening and monitoring efforts, boosts total surplus, and reduces credit risk (i.e., increases the expected time to default). Figure 7 illustrates these findings and shows that they are robust to changes in the $\kappa, \phi, \Lambda,$ and $1/\delta$. Under all parameters considered, bundling increases (initial) monitoring (i.e., $\Delta a_0 > 0$), screening ($\Delta q^* > 0$), and total surplus ($\Delta F_0 > 0$).

Interestingly, whether monitoring and screening task are bundled also has implications for the dynamics of incentives and credit risk. As shown in Section 3, Figure 1, and Corollary 1, incentives to screen and monitor are front-loaded. As a result, they decrease over time when screening and monitoring tasks are bundled implying that monitoring effort decreases over time. In contrast,
Figure 7: The effects of bundling screening and monitoring. $\Delta a_0$ denotes the percentage change in monitoring effort at $t = 0$ caused by bundling. $\Delta q^*$ denotes the percentage change in screening effort $q$ caused by bundling. $\Delta F_{0^-}$ denotes the percentage change in total surplus at $t = 0^-$ caused by bundling. $\Delta \bar{\tau}$ denotes the percentage change in the expected time to default caused by bundling. Outcome variables are plotted as functions of the cost of monitoring $\kappa$, the cost of screening $\phi$, the raw default intensity $\Lambda$, and loan maturity $1/\delta$. 
when screening and monitoring are undertaken by two different entities (i.e., the screener and the monitor), the screener’s incentives (captured by $V_t$) are front-loaded and decrease over time, but the monitor’s incentives $W_t^m$ generally are backloaded and increase over time. The underlying reason is that screener and monitor split the ownership of the pool of loans (before time $\bar{\tau}$): thus, as the screener gradually reduces its exposure to the pool’s performance over time, the monitor’s exposure to the loans and incentives to monitor increase, thus reducing the probability of default (i.e., credit risk). The following Corollary formalizes this result:\(^{18}\)

**Corollary 3** (Dynamics of Incentives with unbundling). Suppose that $\phi > \kappa \bar{q}$ and $W^m(V) = a(V) \phi < F(V)$. Then, under separate tasks, monitoring effort $a(V) = W^m(V)/\phi$ decreases with $V$ and, because $\dot{V}_t < 0$, increases over time. In contrast, under bundled tasks, monitoring effort $a(V)$ increases with $V$ and thus decreases over time when $W(V) < F(V)$ (see Corollary 1).

### 7 Conclusion

We study a dynamic moral hazard problem in which a bank originates a loan to sell it to investors. The bank controls the loan’s default risk through screening at origination and monitoring after origination, both of which are subject to moral hazard. Screening and monitoring incentives are provided by exposing the bank to loan performance. As screening occurs only once at the origination of the loan, incentives are front-loaded and stronger shortly after origination. The optimal contract can be implemented via time-decreasing retention of a stake in the loan so that the bank’s incentives to monitor decrease and credit default risk increases over time. The model implies that there are positive synergies between screening and monitoring incentives, making screening and monitoring complements. Owing to these incentive synergies, screening and monitoring should be carried out by the same entity instead of different entities. In addition, lenders implementing high screening effort (e.g. banks) should also implement high monitoring. The model also shows that by removing moral hazard over screening, credit ratings reduce the size of the stake that the loan originator should retain in the securitized loan and imply that this stake should be constant until the loan matures. A lower stake in securitized loans reduces monitoring incentives so that credit ratings may lead to an increase in default risk.

\(^{18}\)The claim of below Corollary follows from the concavity of the value function (see Proposition 5) and differentiation of $a(V)$ from (43) with respect to $V$.  

---

35
References


Appendix

A Proof of Lemma 1

We first characterize the agent’s monitoring incentives. By the dynamic programming principle and the arguments presented in the main text, the agent chooses monitoring effort \( a_t \) to solve

\[
\max_{a_t \in [0, \bar{a}]} \left( a_t W_t - \frac{\phi a_t^2}{2} \right),
\]

which yields

\[
a_t = \min \left\{ \frac{W_t}{\phi}, \bar{a} \right\}.
\]

Observe that when optimal monitoring effort is interior and \( a_t < \bar{a} \), the above condition simplifies to (7), i.e., \( a_t = \frac{W_t}{\phi} \), which is the first order condition to (A.1). The second order condition to (A.1), i.e., \( \frac{d^2}{da_t^2} \left( a_t W_t - \frac{\phi a_t^2}{2} \right) = -\phi < 0 \), is satisfied.

Second, we characterize the agent’s screening incentives. Note that the agent chooses her screening effort to solve

\[
\max_{q \in [0, \bar{q}]} \left( W_0(q) - \frac{\kappa q^2}{2} \right),
\]

where we make the dependence of \( W_0 \) on \( q \) explicit. Define

\[
V_0(q) = \frac{\partial}{\partial q} W_0(q).
\]

The integral expression (16) and the fact that \( W_t \geq 0 \) (with strict inequality on a set with positive measure) imply that \( V_0(0) > 0 \). Thus, the solution \( q \) to (A.2) satisfies \( q > 0 \).

Now observe that

\[
q = \min \left\{ \frac{V_0(q)}{\kappa}, \bar{q} \right\}
\]

is the unique solution to (A.2) if

\[
\frac{d^2}{dq^2} \left( W_0(q) - \frac{\kappa q^2}{2} \right) = \frac{\partial}{\partial q} V_0(q) - \kappa < 0
\]

holds for any \( q \in [0, \bar{q}] \), in which case the objective in (A.2) is strictly concave over the entire interval [0, \( \bar{q} \)] and the first order approach is valid. When optimal screening effort is interior, condition (A.3) simplifies to (12), i.e., \( q = V_0/\kappa \), which is the first order condition to (A.2).

In what follows, we provide a sufficient condition for (A.4) to hold for all \( q \in [0, \bar{q}] \), which concludes the proof. Define

\[
Y_t(q) = \frac{\partial}{\partial q} V_t(q),
\]

and note that (A.4) can be rewritten as \( Y_0(q) < \kappa \). Next, differentiate (15) with respect to \( q \) to obtain

\[
\frac{dY_t(q)}{dt} = (\gamma + \lambda_t) Y_t(q) - 2V_t(q).
\]
We can integrate the above ODE to obtain

$$Y_t(q) = \int_t^\infty e^{-\gamma(s-t)} - f_t^s \lambda_u du 2V_s(q) ds$$  \hspace{1cm} (A.5)$$

for all $t \geq 0$. In addition, (16) implies

$$V_t(q) = \int_t^\infty e^{-\gamma(s-t)} - f_t^s \lambda_u du W_s(q) ds$$  \hspace{1cm} (A.6)$$

for all $t \geq 0$. Note now that

$$\lambda_t = \Lambda - a_t - q \geq \Lambda - \bar{a} - \bar{q}.$$  \hspace{1cm} (A.7)$$

Next, observe that the agent’s continuation value is bounded from above by

$$W_t \leq F_t = \int_t^\infty e^{-r(s-t)} - f_t^s \lambda_u du \left(1 - \frac{\phi q^2}{2} - (\gamma - r)W_s\right) ds \leq \int_t^\infty e^{-(r+\Lambda-\bar{a}-\bar{q})(s-t)} 1 ds = \frac{1}{r + \Lambda - \bar{a} - \bar{q}} =: W_{max}$$  \hspace{1cm} (A.8)$$

where the first inequality follows from outside investors’ limited liability, i.e., $P_t = F_t - W_t \geq 0$.

Using these two relations (A.7) and (A.8) as well as (A.6), we obtain that

$$V_t(q) < \int_t^\infty e^{-\gamma(s-t)} - f_t^s \lambda_u du W_{max} ds \leq \int_t^\infty e^{-(\gamma+\Lambda-\bar{a}-\bar{q})(s-t)} W_{max} ds$$

$$\leq \frac{1}{\gamma + \Lambda - \bar{a} - \bar{q}} < \frac{1}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})}$$  \hspace{1cm} (A.9)$$

Using this inequality (A.9) and the integral representation in (A.5), we obtain that

$$Y_t(q) = \int_t^\infty e^{-\gamma(s-t)} - f_t^s \lambda_u du 2V_s(q) ds \leq \int_t^\infty e^{-(\gamma+\Lambda-\bar{a}-\bar{q})(s-t)} 2V_s(q) ds$$

$$\leq \frac{2}{(\gamma + \Lambda - \bar{a} - \bar{q})^2(r + \Lambda - \bar{a} - \bar{q})}.$$  \hspace{1cm} (A.10)$$

As a result, a sufficient condition for (A.4), i.e., for

$$Y_0(q) < \kappa$$

to hold for any $q \in [0, \bar{q}]$ is given by

$$\kappa \geq \frac{2}{(\gamma + \Lambda - \bar{a} - \bar{q})^2(r + \Lambda - \bar{a} - \bar{q})}.$$  \hspace{1cm} (A.10)$$

That is, when (A.10) holds, the first order approach is valid and (A.3) or, equivalently, (12) (due to $q < \bar{q}$) pins down screening effort. Note that (A.10) is equivalent to condition (10) (Lemma 1).
B Proof of Proposition 1

To characterize the model solution when screening $q$ is observable and contractible, we proceed in several steps. We first fix $q$ and solve the continuation problem for times $t > 0$. We then determine optimal screening effort, $q = q^B$.

At any time $t > 0$, total surplus, $F_t = P_t + W_t$, can be written as

$$F_t = \int_t^\infty e^{-r(s-t)} - f^*_t \lambda u du (1ds - dC_s) + \int_t^\infty e^{-\gamma(s-t)} - f^*_t \lambda u du \left( dC_s - \frac{\phi a^2}{2} ds \right),$$

where $dC_t$ are the payouts to the agent at time $t$ and $\lambda_t = \Lambda - a_t - q$ is the instantaneous default probability. Note that

$$P_t = \int_t^\infty e^{-r(s-t)} - f^*_t \lambda u du (1ds - dC_s)$$

is the principal’s continuation payoff and

$$W_t = \int_t^\infty e^{-\gamma(s-t)} - f^*_t \lambda u du \left( dC_s - \frac{\phi a^2}{2} ds \right)$$

is the agent’s continuation payoff from time $t$ onward. We can differentiate the expressions for $W_t$ and $P_t$ with respect to time, $t$, to get

$$dP_t = (r + \lambda_t) P_t dt - 1 dt + dC_t$$  \hspace{1cm} (B.11)

$$dW_t = (\gamma + \lambda_t) W_t dt + \frac{\phi a^2}{2} dt - dC_t$$  \hspace{1cm} (B.12)

As a result, the dynamics of total surplus are given by

$$dF_t = dP_t + dW_t$$ \hspace{1cm} (B.13)

$$= (r + \lambda_t) P_t dt - 1 dt + dC_t + (\gamma + \lambda_t) W_t dt - dC_t + \frac{\phi a^2}{2} dt$$

$$= (r + \lambda_t)(P_t + W_t) dt - 1 dt + \frac{\phi a^2}{2} dt - (\gamma - r) W_t dt.$$ \hspace{1cm} (B.14)

We can integrate (B.13) over time, $t$, to get

$$F_t = \int_t^\infty e^{-r(s-t)} - f^*_t \lambda u du \left( 1 - \frac{\phi a^2}{2} (\gamma - r) W_s \right) ds,$$ \hspace{1cm} (B.15)

which is (23) from the main text.

Recall that the agent chooses the payout agreement $C$ to maximize total surplus at time zero

$$F_0 = \frac{\kappa q^2}{2},$$ \hspace{1cm} (B.16)

where $F_0$ is characterized in (B.15). Note that it is always possible to stipulate payouts $dC_t$ to
the agent, which decreases \( W_t \) by amount \( dC_t \). As such, controlling payouts to the agent \( dC_t \) is equivalent to controlling the agent’s continuation payoff \( W_t \). In the following, we take \( W_t \) rather than \( dC_t \) as control variable for the dynamic optimization, and we drop the control variable \( dC_t \).

By the dynamic programming principle, total surplus \( F_t \) must solve at any time \( t > 0 \) the HJB equation
\[
rf_t = \max_{W_t \in [0,F_t], a_t \geq 0} \left( 1 - \frac{\phi a_t^2}{2} - (\gamma - r)W_t + \hat{F}_t - \lambda_tF_t \right),
\]
which is solved subject to the monitoring incentive condition (7) and where \( \hat{F}_t = \frac{dF_t}{dt} \). As default is the only source of uncertainty and as there are no relevant state variables for this dynamic optimization problem, the solution is stationary, so that \( \hat{F}_t = 0 \) and we can omit time sub-scripts (i.e., we write \( F_t = F^B(q) \)). In turn, the HJB equation simplifies to
\[
rf^B(q) = \max_{W \in [0,F^B(q)], a \geq 0} \left( 1 - \frac{\phi a^2}{2} - (\gamma - r)W_t - \lambda F^B(q) \right)
\] (B.17)
subject to the monitoring incentive constraint (7), which can be rearranged to (19).

The maximization in the above HJB equation yields that, if interior, optimal monitoring effort reads
\[
a^B(q) = \frac{F^B(q) - \phi(\gamma - r)}{\phi},
\] (B.18)
and the optimal bank continuation value is \( W^B(q) = \phi a^B(q) \), due to (7). With a slight abuse of notation, if the above expression for \( a^B(q) \) is negative, then optimal monitoring effort \( a^B(q) \) is zero. If the above expression for \( a^B(q) \) exceeds \( \bar{a} \), then optimal monitoring effort \( a^B(q) \) is \( \bar{a} \). Note that the first order condition (B.18) implies \( \phi a^B(q) = W^B(q) < F^B(q) \), so the principal’s limited liability constraint does not bind in optimum.

Optimal monitoring effort implies the instantaneous default probability \( \lambda = \lambda^B(q) = \Lambda - q - a^B(q) \). The law of motion (B.11) and \( dW_t = 0 \) imply then that payouts to the agent take the form \( dC_t = c^B(q)dt \) with
\[
c^B(q) = (\gamma + \lambda^B(q))W^B(q) + \frac{\phi(a^B(q))^2}{2}.
\] (B.19)
That is, payouts to the agent are smooth and positive.

The objective (B.16) can be rewritten as
\[
F^B(q) - \frac{\kappa q^2}{2}.
\] (B.20)
At time \( t = 0 \), the agent chooses screening effort \( q \in [0,\bar{q}] \) to maximize (B.20), so that optimal screening effort \( q^B \) is characterized in (21). As we focus on interior levels, the solution to (21), denoted \( q^B \), is by assumption interior, and therefore satisfies the first order condition \( \frac{\partial F^B(q)}{\partial q} = \kappa q \) for \( q = q^B \).

Finally, we derive the expression (22) for the first order condition to (21). Recall that in optimum (i.e., for \( q = q^B \)), the HJB equation (B.17) holds. Using the envelope theorem, we can differentiate both sides of (B.17) with respect to \( q = q^B \) to obtain under the optimal controls
\((W^B(q), a^B(q))\)

\[(r + \lambda) \frac{\partial F^B(q)}{\partial q} = F^B(q) \iff \frac{\partial F^B(q)}{\partial q} = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} > 0 \quad (B.21)\]

for \(q = q^B\). Utilizing above expression for \(\frac{\partial F^B(q)}{\partial q}\), the first order condition to (21), which is \(\frac{\partial F^B(q)}{\partial q} = \kappa q\), becomes (22), as desired. Also observe that \(\frac{\partial F^B(q)}{\partial q}\) and (B.18) imply that \(a^B(q)\) increases with \(q\).

The second order condition is \(\frac{\partial^2 F^B(q)}{\partial q^2} - \kappa < 0\). The expression in (19) implies that \(F^B(q)\) is convex in \(q\), with \(\frac{\partial^2}{\partial q^2} F^B(q) > 0\). It follows that \(F^B(q) - \frac{\kappa q^2}{2}\) is concave on an interval \([0, q']\) and convex on the interval \([q', \bar{q}]\) for \(q' \leq \bar{q}\). As a result, if there are two values of \(q\), satisfying the first order condition \(\frac{\partial F^B(q)}{\partial q} = \kappa q\), then the smaller one is a maximum and the larger one a local minimum. That is,

\[q^B = \min \left\{ q \in [0, \bar{q}] : \frac{\partial F^B(q)}{\partial q} = \kappa q \right\}. \quad (B.22)\]

As there is — by assumption — an interior solution to (21), above characterization for \(q^B\) is well-defined.

C  Proof of Proposition 2

The proof is split in six parts. Part I characterizes total surplus as a function of the agent’s screening incentives \(V_t = V\) and shows that in optimum, total surplus (i.e., the value function) solves the HJB equation (25). Part II demonstrates that \(\lim_{t \to \infty} V_t = V^B(q)\). Part III characterizes the agent’s initial choice of optimal screening effort \(q^*\). Part IV verifies that \(\kappa q^* < V_0\), and shows that \(\dot{V}_t < 0\) at all times \(t \geq 0\). Part V proves that total surplus (i.e., the value function) decreases in \(V\), is concave when \(F(V) > W(V)\), and satisfies \(\lim_{V \to V^B(q)} F'(V) = 0\). Part VI shows that payouts to the agent are smooth and positive. As stated in the main text, we focus (unless otherwise mentioned) on optimal interior effort levels, \(a_t \in (0, \bar{a})\) and \(q \in (0, \bar{q})\).

We make the following regularity assumption. Throughout, we assume that there exists a unique solution \(F(V)\) to the HJB equation (25) which is continuously differentiable. Further, we assume that the second derivative \(F''(V)\) exists almost everywhere in the state space \((V^B(q), V_0)\) (i.e., the set of points at which \(F'(V)\) is not differentiable is not dense).

C.1  Part I

Our aim is to characterize the model solution when screening effort \(q\) is neither observable nor contractible. As in the proof of Proposition 1, we first fix the choice of \(q\) made at time \(t = 0\) and solve the continuation problem for times \(t > 0\). Recall that according to Lemma 1, the incentive condition (12) holds at time \(t = 0\) so that \(V_0 = \kappa q\).

The agent maximizes total surplus characterized in (B.15):

\[F_t = \int_t^{\infty} e^{-r(s-t)-\int_s^t \lambda_u du} \left( 1 - \frac{\phi a_s^2}{2} - (\gamma - r) W_s \right) ds.\]
Note that it is always possible to stipulate payouts \( dC_t \) to the agent, which decreases \( W_t \) by amount \( dC_t \) and leaves \( V_t \) unchanged. As such, controlling payouts to the agent \( dC_t \) is equivalent to controlling the agent’s continuation payoff \( W_t \). In the following, we take \( W_t \) rather than \( dC_t \) as control variable. Thus, the agent’s optimization problem only depends on the state variable \( V_t \) summarizing the agent’s screening incentives. As a consequence, we can express total surplus as function of \( V_t \), in that \( F_t = F(V_t) \). In what follows, we omit time-subscripts whenever possible.

Recall that screening incentives \( V \) evolve according to (15), i.e.,

\[
\dot{V} = (\gamma + \lambda)V - W.
\]

By the dynamic programming principle, total surplus \( F(V) \) must solve in any state \( V \) the HJB equation

\[
rF(V) = \max_{W \in [0,F(V)],a \in [0,\bar{a}]} \left( 1 - \frac{\phi a^2}{2} - (\gamma - r)W \right) - \lambda F(V) + F'(V)((\gamma + \lambda)V - W),
\]

which is solved subject to the monitoring incentive constraint (7). Recall that both the principal and the agent are subject to limited liability, so that \( W \) and the principal’s payoff \( F(V) - W \) satisfies \( F(V) - W \in [0,F(V)] \) too. The above HJB equation coincides with (25). The maximization in the above HJB equation yields that, if interior, optimal monitoring effort is

\[
a(V) = \frac{F(V) - F'(V)[V + \phi] - (\gamma - r)\phi}{\phi} \land W(C),
\]

which is (28).

Under the benchmark solution from Proposition 1 (for given \( q \)), all model quantities are constant, monitoring is \( a^B(q) \), and the agent’s continuation value is \( W^B(q) = \phi a^B(q) \). As such, screening incentives are constant at level \( V^B(q) \) and by inserting \( \dot{V} = 0 \) and the optimal levels of effort \( a^B(q) \) and continuation value \( W^B(q) = \phi a^B(q) \) into (15), we can solve for

\[
V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}.
\]

It follows that when \( V = V^B(q) \), the continuation surplus is \( F^B(q) \). That is, the surplus function \( F(V) \) satisfies

\[
F(V^B(q)) = F^B(q).
\]

Also note that optimal effort \( a(V) \) satisfies \( a(V^B(q)) = a^B(q) \). In the next Part (i.e., Part II) of the proof, we show that \( \lim_{t \to \infty} V_t = V^B(q) \), which then — together with (C.25) — implies

\[
\lim_{V \to V^B(q)} F(V) = F^B(q),
\]

as well as \( \lim_{V \to V^B(q)} a(V) = a^B(q) \).
As a next step, we prove that \( \lim_{t \to \infty} V_t = V^B(q) \). To do so, we set up the Lagrangian for the total surplus maximization at time \( t = 0 \)

\[
\mathcal{L} = \int_0^\infty e^{-rt} \int_0^\infty \lambda u du \left( 1 - (\gamma - r)W_t - \frac{\phi a_t^2}{2} \right) dt + \ell \left( \kappa q - \int_0^\infty e^{-\gamma t} \lambda u W_t dt \right)
= F_0 + \ell (\kappa q - V_0),
\]

\[\text{(C.26)}\]

where \( \ell \) is the Lagrange multiplier with respect to the screening incentive constraint \( (12) \) and \( W_t = \phi a_t \) is the effort incentive constraint which we directly insert into the objective function.

Next, we rewrite \( (B.13) \) as

\[
dF_t = rF_t dt - 1 dt + (\gamma - r)W_t dt - \frac{\phi a_t^2}{2} dt + \lambda F_t dt,
\]

which can be integrated over time to obtain

\[
F_t = \int_t^\infty \kappa q = \int_t^\infty e^{-\gamma (s-t)} (W_s - \lambda_s V_s) ds.
\]

\[\text{(C.27)}\]

Likewise, we can rewrite \( (15) \) as

\[
dV_t = \gamma V_t dt - W_t dt + \lambda_t V_t dt,
\]

which can be integrated over time to get

\[
V_t = \int_t^\infty e^{-\gamma (s-t)} (W_s - \lambda_s V_s) ds.
\]

\[\text{(C.28)}\]

Using \( (C.27) \) and \( (C.28) \), we can rewrite the Lagrangian \( (C.26) \) as

\[
\mathcal{L} = \int_0^\infty e^{-rt} \left( 1 - (\gamma - r)W_t - \frac{\phi a_t^2}{2} - \lambda_t F_t \right) dt + \ell \left( \kappa q - \int_0^\infty e^{-\gamma t} (W_t - \lambda_t V_t) dt \right),
\]

\[\text{(C.29)}\]

We can maximize the Lagrangian point-wise with respect to \( a_t \), taking into account the monitoring incentive constraint \( (7) \), i.e., \( a_t = W_t/\phi \). If interior, optimal effort \( a_t \) satisfies the first order condition:

\[
e^{-rt} (F_t - (\gamma - r)\phi - \phi a_t) - \ell e^{-\gamma t} (\phi + V_t) = 0
\]

\[\text{(C.30)}\]

Multiplying both sides of \( (C.30) \) by \( e^{rt} \), we obtain

\[
F_t - (\gamma - r)\phi - \phi a_t - \ell e^{-(\gamma - r)t} (\phi + V_t) = 0.
\]

\[\text{(C.31)}\]

We can solve \( (C.31) \) for

\[
a_t = \frac{F_t - (\gamma - r)\phi - \ell e^{-(\gamma - r)t} (V_t + \phi)}{\phi}.
\]

\[\text{(C.32)}\]
Taking the limit $t \to \infty$ in (C.32) leads to
\[
\lim_{t \to \infty} a_t = \lim_{t \to \infty} \left( \frac{F_t - (\gamma - r)\phi}{\phi} \right), \quad (C.33)
\]
as $V_t$ is bounded (see inequality (A.9) in the proof of Lemma 1 and note that by definition, $V_t \geq 0$).

We conjecture (and verify) that, in the limit $t \to \infty$, the solution becomes stationary and $F_t$ and $a_t$ become constant, in that
\[
\lim_{t \to \infty} F_t = \hat{F} \quad \text{and} \quad \lim_{t \to \infty} a_t = \hat{a}
\]
for (endogenous) constants $\hat{F}$ and $\hat{a}$.\(^{19}\) Note that by (C.33),
\[
\hat{a} = \frac{\hat{F} - (\gamma - r)\phi}{\phi} \quad (C.34)
\]
Using that $W_t \to \phi\hat{a}$ and $\lambda_t \to \Lambda - \hat{a} - q$ as $t \to \infty$, we can use (23) to calculate that
\[
\hat{F} = \frac{1 - (\gamma - r)\phi\hat{a} - \frac{\phi\hat{a}^2}{2}}{r + \Lambda - \hat{a} - q}, \quad (C.35)
\]
which confirms that $\lim_{t \to \infty} F_t = \hat{F}$. As
\[
a_t = \arg\max_{a_t \in [0, \hat{a}]} \left( \frac{1 - (\gamma - r)\phi\hat{a} - \frac{\phi\hat{a}^2}{2}}{r + \Lambda - \hat{a} - q} \right), \quad (C.36)
\]
it follows that optimal effort satisfies $\lim_{t \to \infty} a_t = \hat{a}$ for an endogenous constant $\hat{a}$.

Recall the definition of $F^B(q)$ from (B.17). Now note that (C.34) and (C.35) as well as (C.36) jointly imply that $\hat{F} = F^B(q)$ and $\hat{a}^A = a^B(q)$, so that $\hat{W} = W^B(q)$. As a result, it also follows that
\[
\lim_{t \to \infty} V_t = \lim_{t \to \infty} \int_t^\infty e^{-\gamma(s-t)-\int_t^s \alpha_u du} W_s ds = \frac{\phi\hat{a}}{\gamma + \Lambda - \hat{a} - q} = V^B(q) \quad \text{and} \quad \lim_{t \to \infty} \hat{V}_t = 0. \quad (C.37)
\]
As $V_t$ is the only relevant state variable for the dynamic optimization problem, it follows that $V_t$ cannot have a stationary point $V_t \neq V^B(q)$ with $\hat{V}_t = 0$, as otherwise (C.37) would not hold.

That is, when $V_0 = \kappa q > V^B(q)$, it follows that $\hat{V}_t < 0$, with convergence according to (C.37). Likewise, when $V_0 = \kappa q < V^B(q)$, it follows that $\hat{V}_t > 0$, with convergence according to (C.37). In the knife-edge case $V_0 = \kappa q = V^B(q)$, it holds that $V_t = V^B(q)$ and $\hat{V}_t = 0$.

Last, we characterize the limit $\lim_{V \to V^B(q)} F'(V)$. Note that due to (C.25), that is, $F(V^B(q)) = F^B(q)$, and $\lim_{t \to \infty} V_t = V^B(q)$, it follows that $\lim_{V \to V^B(q)} F(V) = F^B(q)$ and $\hat{a}^V = a^B(q)$. We know from Proposition 1 that $W^B(q) < F^B(q)$, so that $\lim_{V \to V^B(q)} W(V) < \lim_{V \to V^B(q)} F(V)$. Thus, for $V$ close to $V^B(q)$, the principal’s limited liability constraint does not bind. Using (C.23),
\[
\lim_{t \to \infty} \hat{F}_t = 0 \quad \text{and} \quad \lim_{t \to \infty} \hat{a}_t = 0.
\]
\(^{19}\)Equivalently,
\[ \lim_{V \to V_B(q)} a(V) = a^B(q) \] becomes equivalent to
\[ \lim_{V \to V_B(q)} F'(V) = 0 \quad (C.38) \]
when \( a^B(q) > 0 \) (which holds by assumption).

As an aside, in the case that \( a^B(q) = V^B(q) = 0 \) — which we do not formally consider — we have
\[ \lim_{V \to V_B(q)} F'(V) = \left( \frac{F^B(q) - (\gamma - r)\phi}{\phi} \right) \leq 0, \quad (C.39) \]
so that \( a(V) \) from (C.23) converges to \( a^B(q) = 0 \) as \( V \to V^B(q) = 0 \).

### C.3 Part III

At time \( t = 0 \), initial screening incentive \( V_0 \) pins down screening effort \( q \) by means of the screening incentive constraint (12). The agent picks the amount of initial screening incentives \( V_0 \) to maximize
\[ F(V_0) - \frac{\kappa q^2}{2} \quad \text{s.t.} \quad V_0 = \kappa q. \quad (C.40) \]
Even if optimal screening is not interior and satisfies \( q^* = \bar{q} \), it would be optimal to set \( V_0 = \kappa q^* \), as \( F(V) \) decreases in \( V > V^B(q) \) and the screening incentive condition (12) is optimally tight.

The first order condition to (C.40) is
\[ \frac{\partial F(V_0)}{\partial q} |_{q=q^*} + F'(V_0)\kappa = \kappa q^*. \quad (C.41) \]

### C.4 Part IV

This part of the proof shows that in optimum (i.e., for \( q = q^* \)), we have \( \kappa q = V_0 > V^B(q) \). Because \( \lim_{t \to \infty} V_t = V^B(q) \) and because there is no stationary point with \( \dot{V}_t = 0 \), \( V_0 > V^B(q) \) implies \( \dot{V}_t < 0 \) at all times \( t \geq 0 \).

Suppose to the contrary that
\[ \kappa q^* = V_0 \leq V^B(q^*) = \frac{W^B(q^*)}{\gamma + \Lambda - a^B(q^*) - q^*}, \quad (C.42) \]
where the last equality follows (C.24). Note that \( W_t \leq F_t \) at all times \( t \geq 0 \) and, in particular, \( W^B(q^*) \leq F^B(q^*) \). We then obtain
\[ \kappa q^* = V_0 \leq \frac{W^B(q^*)}{\gamma + \Lambda - a^B(q^*) - q^*} < \frac{F^B(q^*)}{r + \Lambda - a^B(q^*) - q^*}, \quad (C.43) \]
where the first inequality follows (C.42) and the second inequality uses \( \gamma > r \) and \( W^B(q^*) \leq F^B(q^*) \).

Next, note that (19) implies that \( F^B(q) \) is strictly increasing and convex in \( q \), with \( \frac{\partial^3}{\partial q^3} F^B(q) > \)
Define the function
\[ G(q) = F^B(q) - \frac{\kappa q^2}{2}, \]
which is the objective function in (21). Using (B.21), we obtain
\[ G'(q) = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q. \] (C.44)
We also calculate
\[ G''(q) = \frac{\partial^2}{\partial q^2} F^B(q) - \kappa \quad \text{and} \quad G'''(q) = \frac{\partial^3}{\partial q^3} F^B(q) > 0. \]
Due to \( G'''(q) > 0 \), the function \( G(q) \) is either concave on the entire interval \([0, \bar{q}]\) or concave on an interval \([0, q']\) and convex on the interval \([q', \bar{q}]\) for \( q' < \bar{q} \). This observation implies that \( G(q) \) has at most one local maximum on \([0, \bar{q}]\).

We focus on interior optimal levels of \( q \). Therefore, the maximum of \( G(q) \) on the interval \([0, \bar{q}]\) is denoted by \( q^B \in (0, \bar{q}) \), and satisfies \( G'(q^B) = 0 \) (first order condition) and \( G''(q^B) < 0 \) (second order condition). Thus, \( q^B < \bar{q} \) holds by assumption, and \( q = q^B \) is the unique maximum of \( G(q) \) on \([0, \bar{q}]\). Hence, on \([0, q^B]\), \( G'(q) \neq 0 \), and \( G'(q^B) = 0 \). As \( G''(q^B) < 0 \) and \( G'''(q) > 0 \), it follows that \( G''(q) < 0 \) on the interval \([0, q^B]\). Furthermore, \( G(q) \) must strictly increase on the interval \([0, q^B]\), in that \( G'(q) > 0 \) and \( G''(q) < 0 \) for \( q \in [0, q^B] \).\(^{21}\)

Next, consider the continuous function
\[ K(q) = V^B(q) - \kappa q \] (C.45)
Note that \( a^B(q) \) and \( W^B(q) \) increase with \( q \) (see Proposition 1). Thus, by (C.24), the function \( V^B(q) \) is strictly convex, implying that \( K(q) \) is strictly convex too. Observe that according to (C.43) and (C.44)
\[ K(q) = V^B(q) - \kappa q = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q} - \kappa q < \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q = G'(q). \]
Because i) \( G'(q) \) has a unique root on \([0, q^B]\), ii) because \( K(q) < G'(q) \), iii) because \( K(q) \) is convex, and iv) because \( K(0) > 0, K(q) \) has a unique root \( \hat{q} < q^B \) on \([0, q^B]\) so that \( K(\hat{q}) = 0, K(q) > 0 \) for \( q < \hat{q} \), and \( K(q) < 0 \) for \( q \in (\hat{q}, q^B] \). If \( K(q) \) had a second root \( q_2 \) with \( q^B \geq q_2 > \hat{q} \), then it must be due to convexity that \( K'(q) > 0 \) for \( q \geq q_2 \) and thus \( K(q^B) > G'(q^B) = 0 \), a contradiction.

Next, note that for \( q = \bar{q} \):
\[ K(\bar{q}) = \frac{W^B(\bar{q})}{\gamma + \Lambda - a^B(\bar{q}) - \bar{q}} - \kappa \bar{q} = \frac{a^B(\bar{q})\phi}{\gamma + \Lambda - a^B(\bar{q}) - \bar{q}} - \kappa \bar{q} \leq \frac{\bar{a} \phi}{\gamma + \Lambda - \bar{a} - \bar{q}} - \kappa \bar{q} \leq 0, \]
where the second equality uses (7) and that the incentive constraint for monitoring effort binds, the

\(^{20}\)To see this, note that according to (B.21), \( F^B(q) \) increases with \( q \) and according to Proposition 1, \( a^B(q) \) increases with \( q \) too.

\(^{21}\)As discussed in the proof of Proposition 1, there might be two values of \( q \), satisfying the first order condition \( \kappa q = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} \). In this case, \( q^B \) is the smaller of these two values, and is defined in (B.22).
first inequality uses \( a^B(\hat{q}) \leq \bar{a} \), and the second inequality uses parameter condition (11). Because \( K(q) \) is strictly convex on \([0, \bar{q}]\), \( K(q) \) has precisely one root on \((0, \bar{q})\), which is denoted \( \hat{q} \) and satisfies \( \hat{q} < q^B \). Suppose now \( \kappa q^* = V_0 < V^B(q^*) \), which implies \( K(q^*) > 0 \). Because \( K(q) \) has a unique root on \([0, \bar{q}]\), denoted \( \hat{q} \), it follows that \( q^* < \hat{q} < q^B \).

Total initial surplus can now be written as

\[
F_{0^-} = F_0 - \frac{\kappa(q^*)^2}{2} \leq F^B(q^*) - \frac{\kappa(q^*)^2}{2} < F^B(\hat{q}) - \frac{\kappa(\hat{q})^2}{2},
\]

where the first inequality uses \( F_{0^-} \leq F_B(q) \) (which holds for any \( q \)) and the second inequality uses that \( G(q) = F^B(q) - \frac{\kappa q^2}{2} \) strictly increases on \([0, q^B] \) as well as \( 0 < q^* < \hat{q} < q^B \). As a result, total surplus is higher under a stationary contract that implements screening \( \hat{q} \) and \( V_t = V(\hat{q}) = \kappa \hat{q} \) at all times \( t \geq 0 \), which contradicts the optimality of \( q^* \). Thus, \( V_0 < V^B(q^*) \) cannot be optimal.

Now consider the case \( V_0 = V^B(q^*) = \kappa q^* \), so that \( q^* = \hat{q} < q^B \). Take \( \varepsilon > 0 \) and set \( q^\varepsilon = q^* + \varepsilon \) so that \( q^\varepsilon < q^B \). Because of \( q^* < q^B \), it follows that

\[
\frac{\partial}{\partial q^*} \left( F_B(q^*) - \frac{\kappa(q^*)^2}{2} \right) = G'(q^*) > 0,
\]

which — by (C.46) — exceeds \( F^B(q^*) - \frac{\kappa q^2}{2} \) for \( \varepsilon > 0 \) sufficiently small. The second equality uses that given screening level \( q^\varepsilon \), \( \lim_{V \to V^B(q^\varepsilon)} F'(V) = 0 \) (see (C.38)). However, this contradicts the optimality of \( q = q^* \). Thus, \( V_0 = \kappa q^* > V^B(q^*) \) holds under the optimal choice of \( q = q^* \).

C.5 Part V

In this part, we show \( F'(V) < 0 \) in all accessible states and, in particular, verify our conjecture that \( F'(V_0) \leq 0 \).

First, consider \( F(V) = W(V) \), in that the principal’s limited liability constraint binds. The expression for effort \( a(V) = W(V)/\phi \) in (C.23) implies that \( F'(V) < 0 \), because \( F'(V) \geq 0 \) would imply \( a(V) < F(V)/\phi \) and \( W(V) < F(V) \). Next, suppose that \( F(V) = W(V) \) and insert this relation into the HJB equation (25) to obtain

\[
\gamma F(V) = \max_a \left\{ 1 - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)(\gamma + \lambda)V - F(V) \right\}.
\]

For any points \( V \) at which \( F'(V) \) is differentiable, we can invoke the envelope theorem and differ-
differentiate above ODE with respect

\[ F''(V) = \frac{(F'(V))^2}{(\gamma + \lambda)\dot{V} - F'(V)} = \frac{(F'(V))^2}{\dot{V}} < 0, \]

as we have shown that \( \dot{V} = (\gamma + \lambda)V - W < 0 \) for \( V > V^B(q) \).

Second, suppose that \( F(V) > W(V) \) and the principal’s limited liability constraint does not bind, and consider \( V > V^B(q) \). To start with, note that because the principal’s limited liability constraint does not bind, optimal effort \( a(V) \) solves the first order condition \( \frac{\partial F(V)}{\partial a} = 0 \). For any points \( V \) at which \( F'(V) \) is differentiable, we can then invoke the envelope theorem and totally differentiate the HJB equation (25) under the optimal controls with respect to \( V \) which yields

\[ F''(V) = \frac{-(\gamma - r)F'(V)}{(\gamma + \lambda)V - W}. \tag{C.48} \]

First, note that as shown in Part II of the proof, \( \dot{V} = (\gamma + \lambda)V - W < 0 \) for \( V > V^B(q) \). Thus, \( F''(V) \) has the same sign as \( F'(V) \). It follows by (C.48) that either \( F'(V), F''(V) < 0 \) or \( F'(V), F''(V) \geq 0 \).

Next, let us consider \( V = V^B(q) \) (or the limit \( V \rightarrow V^B(q) \)) so that \( F'(V^B(q)) = 0 \) and — according to the expression for effort (C.23):

\[ a(V^B(q)) = \frac{F(V^B(q)) - (\gamma - r)\phi}{\phi} \Rightarrow W(V^B(q)) < F(V^B(q)), \]

owing to \( \gamma > r \).

If it were \( F'(V), F''(V) \geq 0 \) in a right-neighbourhood of \( V^B(q) \) (i.e., for \( V \in (V^B(q), V^B(q) + \epsilon) \)), then \( F(V) \geq F^B(q) \) for \( V \in (V^B(q), V^B(q) + \epsilon) \). However, it must be that \( F(V) < F^B(q) \) for \( V > V^B(q) \), as providing higher screening incentive \( V > V^B(q) \) than under the benchmark without screening moral hazard for a given level of \( q \) necessarily reduces surplus. As a result, as \( F'(V) \) is continuous, it follows that \( F'(V), F''(V) < 0 \) in a right-neighbourhood of \( V^B(q) \).

Note that when \( F'(V) \) is differentiable, then

\[ \text{sign}(F''(V)) = \begin{cases} -1 & \text{if } W(V) = F(V) \\ \text{sign}(F'(V)) & \text{if } W(V) < F(V). \end{cases} \]

Combined with the fact that \( F'(V), F''(V) < 0 \) in a right-neighbourhood of \( V^B(q) \), it follows that \( F''(V) < 0 \) at all \( V \in (V^B(q), V_0) \) at which \( F'(V) \) is differentiable (and \( F''(V) \) exists). As such, the value function is strictly concave on \( (V^B(q), V_0) \).

### C.6 Part VI

In this part, we show that payouts to the agent are smooth and positive.

We can solve (13) to get the payout rate

\[ c_t = (\gamma + \lambda_t) W_t + \frac{\phi_2 a_t^2}{2} - \dot{W}_t. \tag{C.49} \]
If $F_t = W_t$, note that according to (B.13), $\dot{F}_t = (\gamma + \lambda_t)F_t - 1 + \frac{\phi a_t^2}{2}$. Inserting the law of motion $\dot{F}_t = \dot{W}_t$ into (C.49) yields $c_t = 1 > 0$.

Next, consider $V = V_t$ with $W_t < F_t$. Then, according to (C.23):

$$a(V) = \frac{F(V) - F'(V)[V + \phi] - (\gamma - r)\phi}{\phi},$$

and, provided $a(V)$ is differentiable, then $a'(V) = -\frac{F''(V)[V + \phi]}{\phi} > 0$, as $F''(V) < 0$ when $W < F(V)$. Thus, $\dot{a}_t = a'(V_t)\dot{V}_t < 0$ and, by (7), $\dot{W}_t < 0$. Inserting $\dot{W}_t < 0$ into (C.49) implies $c_t > 0$.

### D Additional Results

#### D.1 Proof of Corollary 1

As the incentive constraint (7) implies $W(V) = \phi a(V)$, it suffices to prove the claims for monitoring effort $a(V)$ for any given $q$. Recall that by (C.23), optimal monitoring effort (if interior) satisfies

$$a(V) = \frac{F(V) - F'(V)[V + \phi] - (\gamma - r)\phi}{\phi},$$

so that (provided that $a(V)$ is differentiable)

$$a'(V) = -\frac{F''(V)[V + \phi]}{\phi}.$$

As $F''(V) < 0$ for $V > V^B(q)$, it follows that $a'(V) > 0$ for $V > V^B(q)$.

Next, note that

$$\lim_{V \to V^B(q)} F'(V) = 0,$$

which implies $\lim_{V \to V^B(q)} a(V) = a^B(q)$.

#### D.2 Proof of Proposition 3 and Details on the Implementation

The proof of Proposition 3 follows from the arguments presented in the main text.

Next, we show how to calculate $\beta_t = \beta(V_t)$, given the optimal contract from Proposition 2 which yields $a(V)$, $W(V) = \phi a(V)$, $c(V)$, and $\dot{V}$ as functions of $V$ as well as optimal screening $q$. Recall that $\lambda_t = \Lambda - a_t - q$, where $a_t = a(V_t)$.

First, observe that

$$D_t = \int_t^\infty e^{-r(s-t)}-\int_t^s \lambda_u du ds,$$

solves the ODE

$$(r + \Lambda - a(V) - q)D(V) = 1 + D'(V)\dot{V}$$

subject to the boundary condition

$$\lim_{V \to V^B(q)} D'(V) = 0 \iff \lim_{V \to V^B(q)} D(V) = \frac{1}{r + \Lambda - a^B(q) - q}.$$
Second, calculate
\[ \dot{W}_t = W'(V_t) \dot{V}_t \quad \text{and} \quad \dot{\beta}(V) = \beta'(V_t) \dot{V}_t, \]
where \( \beta(V) \) is the agent’s retention level in state \( V \) under the proposed implementation of the optimal contract. Third, insert these relations into (32) to obtain the following ODE in state \( V \)
\[ \beta(V) - \beta'(V) \dot{V}D(V) = (\gamma + \Lambda - a(V) - q)W(V) + \frac{\phi a^2}{2} - W'(V)\dot{V}, \quad (D.50) \]
which is solved subject to
\[ \lim_{V \to V^B(q)} \beta'(V) = 0 \iff \lim_{V \to V^B(q)} \beta(V) = c^B(q). \]
Noting there is a one-to-one mapping from time \( t \) to \( V_t = V \), we thus obtain \( \beta_t = \beta(V_t) \) by solving (D.50), as desired. We make the regularity assumption that a unique solution to (D.50) exists.

D.3 Proof of Corollary 2
The proof of corollary 2 follows from Proposition 1 and the arguments presented in the main text. When there is no moral hazard over screening, the agent receives constant payouts \( c^B(q) \), so the optimal contract can be implemented by requiring the agent to retain constant stake \( \beta^B(q) = c^B(q) \) of the pool of loans.

D.4 Proof of Proposition 4
Analogous to the solution of the baseline, we first provide the solution to the continuation problem for \( t \geq 0 \) and a given level of \( q \). Then, we determine the optimal screening level \( q \), taking into account the solution to the continuation problem.

We characterize the model solution when there is no moral hazard over monitoring, so that the incentive constraint (7) does not apply. As in the baseline, the agent maximizes total surplus at time \( t = 0 \) and screening incentives \( V \) is the only relevant state variable for the dynamic optimization, which follows (15). The agent’s continuation payoff follows (13). As such, total surplus (i.e., the value function) is a function of \( V \) only and solves the HJB equation
\[ rF(V) = \max_{W \in [0,F(V)],a \in [0,\bar{a}]} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)((\gamma + \lambda)V - W) \right\}. \quad (D.51) \]
The maximization with respect to monitoring effort, \( a \), yields that, if interior, optimal monitoring effort is
\[ a(V) = \frac{F(V) - F'(V)\dot{V}}{\phi}. \]
The maximization with respect to the agent’s deferred compensation yields that
\[ W(V) \begin{cases} = 0 & \text{if } F'(V) > -(\gamma - r) \\ \in [0,F(V)] & \text{if } F'(V) = -(\gamma - r) \\ = F(V) & \text{if } F'(V) < -(\gamma - r). \end{cases} \quad (D.52) \]
Note now that when screening is observable and contractible (in addition to monitoring being observable and contractible), then \( V^B(q) = W^B(q) = 0 \). As in the baseline, it follows that \( \lim_{t \to \infty} V_t = V^B(q) = 0 \). As a result, it must be that \( V_t < 0 \) at all times \( t \geq 0 \), in that
\[
\dot{V} = (\gamma + \lambda)V - W(V) < 0.
\]

Owing to (D.52), this requires that \( W(V) > 0 \) for \( V > 0 \) and therefore \( F'(V) \leq -(\gamma - r) \) for \( V > 0 \), owing to (D.52).

Thus, it is (at least) weakly optimal to stipulate \( W(V) = F(V) \), which we can insert into the HJB equation (D.51) to obtain
\[
\gamma F(V) = \max_{a \in [0, \bar{a}]} \left\{ 1 - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)((\gamma + \lambda) V - F(V)) \right\}.
\]

Let us assume that \( F''(V) \) exists and is well-defined. Using the envelope theorem, we totally differentiate the HJB equation (D.53) (under the optimal control \( a = a(V) \)) with respect to \( V \), which yields
\[
F''(V) = \frac{(F'(V))^2}{(\gamma + \lambda)V - F(V)}.
\]

Due to \( \dot{V} = (\gamma + \lambda)V - F(V) < 0 \), we have \( F''(V) < 0 \), i.e., \( F(V) \) is strictly concave. That is, \( F(V) \) is strictly concave for \( V > 0 \). If there exists now \( \dot{V} > 0 \) with \( F'(V) = -(\gamma - r) \), then there exists \( 0 < V' < \dot{V} \) with \( F'(V') > -(\gamma - r) \), a contradiction. As a result, \( F'(V') < -(\gamma - r) \) for all \( V > 0 \).

As \( V \) approaches zero, it must be that \( \dot{V} \) approaches zero too, as — by definition — \( V \) cannot become negative. As such, \( W(0) = 0 \), which requires by means of (D.52) that \( F'(0) \geq -(\gamma - r) \). As \( F'(V) < -(\gamma - r) \) and \( F''(V) \) is continuous for all \( V > 0 \), it follows that \( F'(0) = -(\gamma - r) \).

As in the baseline, optimal screening effort \( q^* \) maximizes total initial surplus \( F_0 = F(V_0) - \frac{\kappa q^2}{2} \) subject to the incentive constraint \( V_0 = \kappa q \).

**Implementation of the Optimal Contract**

We are now in the position to characterize the implementation of the optimal contract, described above. For this sake, note that one unit claim in the pool of loans has a payout rate 1.

Next, we characterize the payouts to the agent and, doing so, we omit time subscripts unless confusion is likely to arise. Using the law of motion for the agent’s continuation payoff
\[
\frac{dW}{dt} = (\gamma + \lambda)W + \frac{\phi a^2}{2}dt - dC,
\]
it follows that the agent receives a payout \( dC = F(0) \) at the time \( V \) reaches zero. When \( V > 0 \), then \( F(V) = W(V) \), and according to (B.13) for \( W(V) = F(V) \):
\[
\frac{dW}{dt} = (\gamma + \lambda)Wdt + \frac{\phi a^2}{2}dt - 1dt = dF_t,
\]
yielding
\[
dC = 1dt,
\]
which equals coupon payments over an instant $dt$.

As a result, the contract is implemented by requiring the agent to fully retain the pool of loans until time $\tilde{\tau} = \inf\{ t \geq 0 : V_t = 0 \}$ and to sell them to outside investors at the time $V$ reaches zero. When $V = 0$ at time $\tilde{\tau}$, the agent sells her entire stake to the principal (outside investors), and she receives the fair price of $F(0)$ dollars, implementing the desired payout $dC = F(0)$ to the agent.

### D.5 Solution with Separation of Screening and Monitoring — Proof of Proposition 5

Analogous to the solution of the baseline, we first provide the solution to the continuation problem for $t \geq 0$ and a given level of $q$. Then, we determine the optimal screening level $q$, taking into account the solution to the continuation problem.

Denote payouts to the screener as $dC^s_t$ and payouts to the monitor as $dC^m_t$. The contracts to screener and monitor stipulate payouts $\{dC^s_t\}$ and $\{dC^m_t\}$ to screener and monitor respectively, and are chosen to maximize total surplus. Recall that in all other aspects, the screener and monitor are symmetric and have the same preferences, an assumption that facilitates maximal comparability to the baseline.

Define the screener’s continuation value (from time $t$ onward) as

$$W^s_t = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} dC^s_s$$

and the monitor’s continuation value (from time $t$ onward) as

$$W^m_t = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} \left( dC^m_s - \frac{\phi a^2}{2} ds \right),$$

where $a_t$ is monitoring effort and $q$ is screening effort, leading to $\lambda_t = \Lambda - a_t - q$. As such, we obtain

$$dW^s_t = (\gamma + \lambda_t)W^s_t dt - dC^s_t \tag{D.54}$$

$$dW^m_t = (\gamma + \lambda_t)W^m_t dt - dC^m_t + \frac{\phi a^2}{2} dt \tag{D.55}$$

As $dC^s_t$ and $dC^m_t$ are not sign-restricted, we can treat $W^s_t$ and $W^m_t$ as control variables in the dynamic optimization problem, while dropping the controls $dC^s_t$ and $dC^m_t$.

Analogous to the baseline and the model extension from Section 4, we define the screener’s screening incentives at time $t$ as

$$V_t = \int_t^\infty e^{-(\gamma+\delta)(s-t)-\int_t^s \lambda_u du} W^s_s ds, \tag{D.56}$$

where $\delta \geq 0$ is debt maturity (the baseline is obtained for $\delta = 0$). Thus,

$$dV_t = (\gamma + \lambda_t + \delta)V_t dt - W^s_t dt. \tag{D.57}$$
As in the baseline version of the model, optimal screening is pinned down by

\[ V_0 = \kappa q, \]

which is analogous to (12). Optimal monitoring (if interior) is pinned down by

\[ a_t = \frac{W^m_t}{\phi}, \]

which is analogous to (7).

The optimal contracts to both the screener and monitor are designed to dynamically maximize total surplus \( F_t \). Total surplus \( F_t \) can be rewritten (using arguments analogous to the ones that lead to (B.15)) as

\[
F_t = \int_t^\infty e^{-(r+\delta)(s-t)} f_s \lambda_s du \left( 1 - \frac{\phi a^2}{2} - (\gamma - r)(W^s_s + W^m_s) + \delta F^*_s \right) ds,
\]

where \( F^*_s \) is the (continuation) surplus “just after” maturity (which occurs at rate \( \delta \)). We will specify the exact form of \( F^*_s \) below.

As in the baseline version of the model, screening incentives \( V \) is the only state variable for the dynamic optimization problem, while \( W^m \) and \( W^s \) can be treated as control variables. Accordingly, by the dynamic programming principle, total surplus \( F(V) \) solves the HJB equation

\[
(r + \delta)F(V) = \max_{a, W^m, W^s} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)(W^m + W^s) - \lambda F(V) + \delta F^* + F'(V)((\gamma + \lambda + \delta)V - W^s) \right\},
\]

which is (40). Note that limited liability requires that \( W^m \in [0, F(V) - W^s] \) and \( W^s \in [0, F(V) - W^m] \) and incentive compatibility with respect to monitoring requires that \( W^m = a \phi \).

The maximization with respect to the screener’s deferred compensation \( W^s \) yields that

\[
W^s(V) = \begin{cases} 
0 & \text{if } F'(V) > - (\gamma - r) \\
[0, F(V) - W^m(V)] & \text{if } F'(V) = - (\gamma - r) \\
F(V) - W^m(V) & \text{if } F'(V) < - (\gamma - r).
\end{cases}
\]

As in the baseline, it follows that \( \lim_{t \to \infty} V_t = V^B(q) \), where \( V^B(q) \) is the level of screening incentives in the benchmark without screening moral hazard (given \( q \)). It follows that \( V^B(q) = 0 \), as absent screening moral hazard it is optimal to set \( V_t = W^s_t = 0 \) at all times \( t \geq 0 \).

As a result, it must be that \( \dot{V}_t < 0 \) at all times \( t \geq 0 \), in that

\[
\dot{V} = (\gamma + \lambda + \delta)V - W^s(V) < 0.
\]

Owing to (D.60), this requires that \( W^s(V) > 0 \) for \( V > 0 \) and therefore \( F'(V) \leq - (\gamma - r) \) for \( V > 0 \). Next, suppose that \( F'(V) < - (\gamma - r) \) for \( V > 0 \), so \( W^s(V) = F(V) - W^m(V) \). Inserting
this expression into (D.59) and simplifying leads to the ordinary differential equation

\[(\gamma + \delta)F(V) = \max_{a,W^m} \left\{1 - \frac{\phi a^2}{2} - \lambda F(V) + \delta F'(V))((\gamma + \lambda + \delta)V - F(V) + W^m)\right\}, \quad (D.61)\]

whereby \(a = W^m/\phi\).

As in the main text (compare Section 4), we consider \(F^\delta = F(V)\), so (D.61) simplifies to

\[
\gamma F(V) = \max_{a,W^m} \left\{1 - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)((\gamma + \lambda + \delta)V - F(V) + W^m)\right\}. \quad (D.62)
\]

Using the envelope theorem to totally differentiate the HJB equation (D.62) (under the optimal controls) with respect to \(V\) yields

\[
F''(V) = \frac{(F'(V))^2 - \delta F'(V)}{(\gamma + \lambda + \delta)V - F(V) + W^m} = \frac{(F'(V))^2 - \delta F'(V)}{V},
\]

where the second equality uses \(W^s(V) = F(V) - W^m(V)\) and \(\dot{V} = (\gamma + \lambda + \delta)V - F(V) + W^m\) (see (D.57)). It must be that \(F'(V) < 0\) for \(V > 0\), as otherwise there exists a point \(V' > 0\) with \(F(V') > F^B(q)\) which cannot be. That is, \(F(V)\) is strictly concave for \(V > 0\). If there exists now \(\dot{V} > 0\) with \(F'(\dot{V}) = -(\gamma - r)\), then there exists \(0 < V' < \dot{V}\) with \(F'(V') > -(\gamma - r)\), a contradiction. As a result, \(F'(V) < -(\gamma - r)\) for all \(V > 0\).

The maximization in (D.61) with respect to monitoring effort yields

\[
a(V) = \frac{F(V) - F'(V)V + F'(V)\phi}{\phi}, \quad (D.63)
\]

which is (43). When \(V\) approaches zero, it must be that \(\dot{V}\) approaches zero too, as — by definition — \(V\) cannot become negative. As such, \(W^s(0)\) approaches zero, which requires by means of (D.60) that \(F'(0) \geq -(\gamma - r)\). As \(F'(V) < -(\gamma - r)\) for all \(V > 0\), it follows — by continuity of \(F'(V)\) — that \(\lim_{V \to 0} F'(V) = -(\gamma - r)\). An alternative way to derive this boundary condition is as follows. Comparing (19) with (D.61), one can see that

\[
\lim_{V \to V_0} F(V) = F^B(q)
\]

is equivalent to

\[
\lim_{V \to 0} F'(V) = -(\gamma - r),
\]

which is then natural the boundary condition for the ODE (D.61) as \(V\) approaches zero.

Because

\[
dW^s_t = (\gamma + \lambda_t)W^s_t dt - dC^s_t,
\]

the screener receives a payout of

\[
dC^s = W^s(0) = F(0) - W^m(0)
\]
dollars at the time \(V\) reaches zero.
As in the baseline, optimal screening effort $q^*$ maximizes total initial surplus $F_0^- = F(V_0) - \frac{\kappa q^2}{2}$ subject to the incentive constraint $V_0 = \kappa q$. 