Screening and Monitoring Corporate Loans*

Sebastian Gryglewicz† Simon Mayer‡ Erwan Morellec§

November 15, 2021

Abstract

We study a dynamic moral hazard problem in which a bank originates a pool of loans that it sells to competitive investors via securitization. The bank controls the default risk of the loans by screening at origination and monitoring after origination, but it is subject to moral hazard. The optimal contract between the bank and investors can be implemented via a time-decreasing stake within the loan pool, so that the bank's monitoring incentives decrease and default risk increases over time. Screening and monitoring have positive incentive synergies and are complements, rendering bundling optimal. Credit ratings distort incentives, potentially increasing credit risk.

*We would like to thank Doug Diamond, Barney Hartman-Glaser, Denis Gromb, and Gustavo Manso for comments. Erwan Morellec acknowledges financial support from the Swiss Finance Institute.

†Erasmus University Rotterdam. Email: gryglewicz@ese.eur.nl

‡University of Chicago, Booth School of Business, and HEC Paris. E-mail: simon.mayer@chicagobooth.edu.

§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, in part due to the growth of leveraged loans. As discussed in Cordell, Roberts, and Schwert (2021), "two-thirds, or $2.1 trillion, of leveraged loan issuance since the 2008 financial crisis has been funded by collateralized loan obligations" (CLOs) in which a broad array of financial institutions, such as pension funds, mutual funds, or insurers, invest.[1] 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 mortgage and corporate debt markets 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 through securitization (Drucker and Puri (2009)), thereby reducing or eliminating its exposure to default risk, it may not have sufficient incentives to screen and monitor borrowers (Pennacchi (1988) or Gorton and Pennacchi (1995)).[2] Indeed, monitoring and screening borrowers is costly, but investors generally do not observe the quality of the loans they invest in. This creates a complex agency problem between the loan originator and investors that extends over the entire life of the loan. As screening occurs only once at origination while monitoring occurs after origination, moral hazard over screening and monitoring likely have different implications for incentive provision and credit risk. The objective of this paper is to study optimal incentive provision for screening and monitoring in the context of credit securitization and to derive implications for regulation, the value of credit ratings, and their effects on credit risk.

To do so, we develop a dynamic agency model in which a bank (the agent) originates a pool of loans and sells these loans to competitive investors (the principal). The pool of loans generates coupon payments at a constant rate up to default (or maturity). When originating the loans, the bank may undertake a costly screening effort that results in lower expected default rates. It can also continuously monitor the loans at a cost afterward to further control default risk. If the bank decides to shirk, the loans will have a higher default rate. Screening at origination and monitoring after origination are not observable by investors, leading to moral hazard. The bank has a lower valuation for the loans than investors due to a higher discount rate (arising, e.g., from regulatory constraints). There are therefore gains from selling the loans to investors, which is done through securitization. Securitizing loans reduces the bank's exposure to loan performance and undermines its incentives, thereby increasing credit risk and decreasing the value of securitized loans.

[1] CLOs 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.

[2] As banks have shifted from an "originate-to-retain" model to "originate-to-distribute," their direct exposures to credit risk have indeed declined over time.

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 agent to loan performance. As the agent is protected by limited liability, this is achieved by delaying her payouts so that she loses her expected future payments upon default. However, delaying payments is costly due to the agent's impatience. 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 fraction of the loans at origination and gradually sells these loans over time.

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 bank) to retain a stake in the pool of loans that it gradually sells to investors.

The model generates several novel implications for the relation between screening and monitoring in credit securitization and their relation to credit quality. In particular, we 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 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 (e.g., via loan covenants) typically finance high quality borrowers with high priority loans.[3] 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, 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.

Our model also allows us to examine the effects of credit ratings on the incentives of loan originators and credit risk. 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." In effect, 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 less incentives for the agent through delayed payments. Therefore, credit ratings weaken monitoring incentives and have an ambiguous impact on credit risk. Our model predicts that credit ratings generate value and are useful in alleviating credit risk for firms that implement long-term financing by rolling over short-maturity debt, and for firms with a high cost of screening due to, e.g., low asset tangibility.

Next, we study how debt maturity affects the bank's incentives to screen and monitor loans, and vice versa. To compare loans of different maturity, we consider a setting in which loans are repeatedly refinanced and securitized at issuance. That is, when existing loans mature, borrowers refinance by issuing identical loans and the bank (originator) writes a new securitization contract with outside investors that runs until the next maturity date. As such, the setting with finite debt maturity features a repeated contracting relationship between the bank and outside investors. At origination, the bank cannot commit to future contracts with outside investors. As screening only occurs at origination, the choice of screening is considered sunk when the loans are refinanced *after* origination and the bank and outside investors write a new contract. Thus, any contract written after origination provides relatively low screening incentives so that short loan maturity undermines

[3]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.

screening incentives. To counteract the adverse effect of short loan maturity on screening, the bank maintains a larger exposure to the securitized loans at origination which improves monitoring incentives. That is, short maturity debt features less screening and more monitoring which, as we show in the paper, implies that credit risk is higher for firms that implement long-term financing by rolling over short-maturity debt. A direct consequence of this result is that credit ratings are particularly valuable for these firms, as short maturity loans are associated with low screening and credit ratings alleviate moral hazard over screening.

In some applications of credit securitization, such as mortgage lending, screening and monitoring of loans are often undertaken by separate entities: an originator responsible for screening and a servicing company in charge of monitoring (Demiroglu and James (2012)). In other settings, the two tasks 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 pool of securitized loans. However, raising one agent's incentives and stake in the pool of loans 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 tasks, so 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.

This paper builds on the literature that studies dynamic contracts in continuous time, starting with DeMarzo and Sannikov (2006), Biais, Mariotti, Plantin, and Rochet (2007), and Sannikov (2008). Recent contributions include He (2009, 2011, 2012), Biais, Mariotti, Rochet, and Villeneuve (2010), DeMarzo, Fishman, He, and Wang (2012), Green and Taylor (2016), Varas (2018), Marinovic and Varas (2019), Gryglewicz, Mayer, and Morellec (2020), Mayer (2020), Feng and Westerfield (2021), and Feng, Taylor, Westerfield, and Zhang (2021). Within this rapidly growing literature, Piskorski and Westerfield (2016), Malenko (2019), and Gryglewicz and Mayer (2021) analyze incentive provision with optimal dynamic contracts and monitoring. Likewise, Halac and Prat (2016), Varas, Marinovic, and Skrzypacz (2020), and Marinovic and Szydlowski (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 paper that is closely related to our work, Hartman-Glaser, Piskorski, and Tchistyi (2012) study optimal securitization and screening of loans under moral hazard. In their model, the optimal contract features a single payout to the agent when sufficient time has elapsed after loan origination and the loans have not defaulted yet. Importantly, they demonstrate that while the 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. Hoffmann, Inderst, and Opp (2020) analyze optimal regulation in a setup similar to that in Hartman-Glaser et al. (2012). Unlike ours, these papers do not model screening and monitoring and, as a result, cannot study optimal 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, e.g., in 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.

Our paper is also related to papers that study bank screening and monitoring in static settings, such as Diamond (1984), Gorton and Pennacchi (1995), or Holmstrom (1989), and to papers that study optimal credit securitization in static settings, such as DeMarzo and Duffie (1999) and more recently Daley et al. (2020). Because these models are static, they do not explicitly distinguish between monitoring after loan origination and screening of loans at origination.

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 4 and

5 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 summarizes our empirical predictions. Section 8 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 pool of loans that can be sold to competitive outside investors (the principal or "they") immediately after origination. The pool of loans promises a constant flow payoff (coupon payments) $\mu > 0$ up to its default, which occurs at the random time τ. The default time τ arrives according to a jump process $dN_t \in \{0, 1\}$ with 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 pool of loans defaults with probability $\mathbb{E}dN_t = \lambda_t dt$. The default rate λ_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

$$\lambda_t = \Lambda - a_t - q. \tag{1}$$

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 λ_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

$$\bar{\tau}_t = \int_t^{\infty} e^{-\int_t^s \lambda_u du} ds, \tag{2}$$

which reflects credit quality and/or credit risk. 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 $\bar{\tau} := \bar{\tau}_0$ corresponds to high credit risk.

Screening entails a cost $\frac{1}{2}\kappa q^2$ to the bank at time zero. Monitoring entails a flow cost $\frac{1}{2}\phi a_t^2$ 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 ex-ante 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 λ_t in a symmetric and independent way. If the bank decides to shirk on either task, the loans will have a higher default rate.

Both the principal (investors) and the agent (the bank) are risk neutral. The principal discounts cash flows at rate $r > 0$. The agent is more impatient and applies a discount 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). Alternatively, the discount rate differential can account for 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 loans—or a security whose payoff depends on loan performance—to outside investors, a process that we term securitization. Securitization works as follows. At inception, the bank designs a financial contract or, equivalently, a security $\mathcal{C}$ that is sold to competitive investors at price P_0. The contract $\mathcal{C} = \{dC_t, \hat{a}_t, \hat{q}\}$ represents a claim on the pool of loans originated by the bank and stipulates a profit-sharing rule C of the overall loan payments μdt, so that the bank receives dC_t and the investors receive $\mu dt - dC_t$ dollars over each time interval $[t, t + dt]$. The contract $\mathcal{C}$ also stipulates contracted/proposed 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 $\mathcal{C}$ 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 $dC_t \in [0, \mu dt]$, so the bank receives positive payouts $dC_t \geq 0$ over each time interval $[t, t + dt]$. Conversely, under the optimal security (contract) $\mathcal{C}$, investors receive $\mu dt - dC_t \geq 0$.

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 inception at time $t = 0^-$, the principal and the agent sign a contract $\mathcal{C}$, after which the agent chooses her screening effort q. Notably, given the contract $\mathcal{C}$, the agent chooses screening effort q and monitoring effort $\{a_t\}$ to maximize the expected present value of private profits

$$W_{0^-} = \max_{q, \{a_t\}} \mathbb{E}\left[\int_0^\infty e^{-\gamma t}\left(dC_t - \frac{\phi a_t^2}{2} dt\right)\right] - \frac{\kappa q^2}{2}, \tag{3}$$

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 in the optimal contract, the bank should not be rewarded for default because this outcome indicates either poor monitoring, poor screening, or both. Hence, no positive payments should be made to the bank after time τ; 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 all $t \geq \tau$. Altogether, we thus have that the payments to the bank satisfy $dC_t = 0$ for $t \geq \tau$. We additionally conjecture (and later verify) that after securitization at time $t = 0^-$, payouts to the bank are smooth in that $dC_t = c_t dt$ for a flow compensation stream c_t at time $t \geq 0$.

The price that outside investors pay for a contract $\mathcal{C}$ at time $t = 0^-$ is given by $P_{0^-} = P_0$ where the time-t price of the security is

$$P_t = \mathbb{E}_t \left[\int_t^\tau e^{-r(s-t)} (\mu - c_s) ds \right] = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} (\mu - c_s) ds. \tag{4}$$

In equation (4), the second equality integrates the default intensity λ_s over the relevant time interval. The bank receives P_0 dollars at time $t = 0^-$ from the securitization process, i.e., from selling the security to investors, in that $dC_{0^-} = P_0$.

As outside investors are competitive, the bank chooses the security that maximizes total initial surplus $F_{0^-} := W_{0^-} + P_0$ at time $t = 0^-$. That is, the bank solves

$$\max_{\mathcal{C}} F_{0^-}, \tag{5}$$

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

Under the contract $\mathcal{C}$, the agent's continuation payoff at time $t \geq 0$ is

$$W_t := \mathbb{E} \left[\int_t^\tau e^{-\gamma(s-t)} \left(c_s - \frac{\phi a_s^2}{2} \right) ds \right] = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} \left(c_s - \frac{\phi a_s^2}{2} \right) ds, \tag{6}$$

where the second equality integrates the default intensity λ_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 interior monitoring effort $a_t \in (0, \bar{a})$, the bank's optimal monitoring effort is

$$a_t = \frac{W_t}{\phi}. \tag{7}$$

Equation (7) describes the incentive constraint for monitoring effort, in that incentive compatibility requires $\hat{a}_t = a_t = \frac{W_t}{\phi}$ 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.

While monitoring a_t impacts the instantaneous default intensity λ_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, $\frac{\kappa q^2}{2}$:

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

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

$$\frac{\partial}{\partial q} W_0(q) = \kappa q. \tag{9}$$

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*

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

Incentive conditions (7) and (12) hold and uniquely pin down the agent's monitoring and screening effort. The incentive conditions (7) and (12) 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:

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

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

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

$$q = \frac{V_0}{\kappa}. \tag{12}$$

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 requires $q = \hat{q} = \frac{V_0}{\kappa}$.

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[4]

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

[4] For a general payout process dC_t, it follows similarly that $dW_t = (\gamma + \lambda_t) W_t dt + \frac{\phi a_t^2}{2} dt - dC_t$.

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:

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

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[5]

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

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

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

Expression (16) shows that screening incentives V_t decrease with the loan default rate. As a result, they are stronger if the loans are 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 the performance of the pool of loans 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.

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

[5] The first order condition with respect to monitoring, $\frac{W_t}{\partial a_t} = 0$, so that by the envelope theorem, $\frac{\partial}{\partial q}\frac{W_t}{\partial a_t} = 0$.

$c_t = c > 0$ 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

$$c = (\gamma + \Lambda - a - q)W + \frac{\phi a^2}{2}. \tag{17}$$

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:

$$P^B(q) = \frac{\mu - c}{r + \Lambda - a - q}. \tag{18}$$

$P^B(q)$ is the discounted stream of flow payouts to outside investors, $\mu - c$, 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

$$F^B(q) = \max_{W \in [0, F^B(q)]} \left(\underbrace{\frac{\mu}{r + \Lambda - a - q}}_{\text{Market value}} - \underbrace{\frac{(\gamma - r)W}{r + \Lambda - a - q}}_{\text{Agency cost}} - \underbrace{\frac{\frac{\phi a^2}{2}}{r + \Lambda - a - q}}_{\text{Monitoring cost}} \right), \tag{19}$$

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

$$V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}. \tag{20}$$

Equation (20) characterizes the agent's screening incentives under best 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:

$$q^B = \arg\max_{q\in[0,\bar{q}]} \Bigg(\underbrace{F^B(q)}_{\substack{\text{Surplus after}\\ \text{screening}}} - \underbrace{\frac{\kappa q^2}{2}}_{\substack{\text{Screening}\\ \text{cost}}} \Bigg) \tag{21}$$

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

$$\kappa q^B = \frac{F^B(q^B)}{r + \Lambda - a^B(q^B) - q^B}. \tag{22}$$

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:[6]

$$\begin{aligned} F_t &= \underbrace{\int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}(\mu - c_s)ds}_{=P_t} + \underbrace{\int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du}\left(c_s - \frac{\phi a_s^2}{2}\right) ds}_{=W_t} \\ &= \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}\left(\mu - \frac{\phi a_s^2}{2} - (\gamma - r)W_s\right) ds. \end{aligned} \tag{23}$$

[6]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

$$\dot{F}_t = (r+\lambda_t)P_t - \mu + c_t + (\gamma+\lambda_t)W_t - c_t + \frac{\phi a_t^2}{2} = (r+\Lambda_t)(\underbrace{P_t + W_t}_{=F_t}) - \mu + \frac{\phi a_t^2}{2} - (\gamma - r)W_t.$$

The above expression can be integrated over time, t, to arrive at the second line of (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 τ, 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:[7]

$$\begin{aligned} rF(V,W) = \max_{a,c} \Bigg\{ &\mu - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V,W) \\ &+ F_V(V,W)((\gamma+\lambda)V - W) + F_W(V,W)\left((\gamma+\lambda)W + \frac{\phi a^2}{2} - c\right)\Bigg\}, \end{aligned} \tag{24}$$

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 $dC = cdt$. Note that it is always possible to stipulate that the bank receives a payout of Δ dollars, which leaves V unchanged but changes W by $-\Delta$ dollars.[8] 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

$$rF(V) = \max_{a,W} \left\{ \mu - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)\big((\gamma+\lambda)V - W\big) \right\}, \tag{25}$$

where 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

[7] For a derivation conjecture that $F_t = F(V_t, W_t)$, so $\dot{F}_t = F_V(V_t, W_t)\dot{V}_t + F_W(V_t, W_t)\dot{W}_t$. Differentiate (23) with respect to time to get

$$\dot{F}_t = (r + \lambda_t)F_t - \mu + \frac{\phi a_t^2}{2} - (\gamma - r)W_t,$$

which becomes (24) after inserting $\dot{F}_t = F_V(V_t, W_t)\dot{V}_t + F_W(V_t, W_t)\dot{W}_t$ and $F_t = F(V_t, W_t)$.

[8] If payouts to the bank are not smooth, then it follows similar to (13) that

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

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

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 $t = 0^-$, 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 ∞, 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

$$\lim_{V \to V^B(q)} F(V) = F^B(q). \tag{26}$$

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 when $F(V) > W(V)$ and the principal's limited liability constraint does not bind, in that $F''(V) < 0$.

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 the ODE (25) subject to condition (26). The value $V^B = V^B(q)$ is given by $V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}$.*

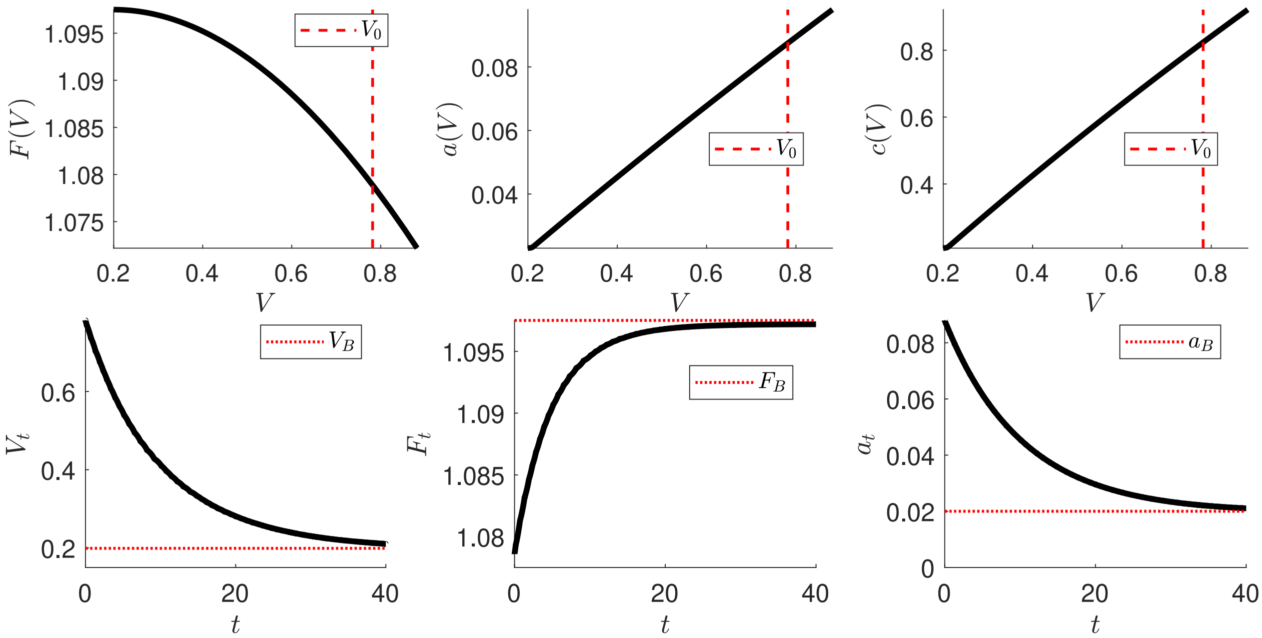

Figure 1: **Optimal contract**. In the upper panels, the dashed red line denotes the V_0. In the lower panels, the dotted red line denotes the benchmark levels that are attained in the limit $t \to \infty$.

2. *Optimal monitoring is characterized by the maximization in the HJB equation* (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* $F'(V^B(q)) = 0$*, so that* $F'(V) < 0$ *for* $V > V^B(q)$*. Whenever* $F(V) > W(V)$*, the value function is strictly concave.*

6. *Payouts to the agent are smooth and positive.*

Figure 1 provides a numerical example of the optimal contract under our baseline parameters. For the numerical analysis, we normalize $r = 0$, $\mu = 1$, 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 pool of loans has value $\mu/\Lambda = 1$. In addition, we set $\gamma = 0.1$ and $\phi = \kappa = 9$ to generate the desired trade-offs. Last, we pick $\bar{a} = 0.125$ and $\bar{q} = 0.2$ to satisfy conditions (10) and (11). Our parameter choice also implies that screening and monitoring effort are interior at all times (i.e., the constraints $a_t \leq a$ and $q \leq \bar{q}$ never bind).

The three upper panels of Figure 1 plot the total surplus $F(V)$, monitoring $a(V)$, and the agent's flow payouts $c(V)$ to the agent as functions of the state variable V. Observe that flow payouts to the agent are always positive. Likewise, as $c_t < \mu = 1$ 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 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 the screening task at time $t = 0$. 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. Intuitively, the provision of screening incentives distorts monitoring incentives upward, which is costly and curbs total surplus. Over time, these distortions and screening incentives 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 effort follows from the first-order condition of the Hamilton-Jacobi-Bellman equation (25):

$$a(V) = \frac{\overbrace{F(V)}^{\substack{\text{Reduction of}\\ \text{default risk}}} \overbrace{-F'(V)(V+\phi)}^{\substack{\text{Screening}\\ \text{incentives}(>0)}} - \overbrace{(\gamma - r)\phi}^{\substack{\text{Agency}\\ \text{costs}}}}{\underbrace{\phi}_{\text{Physical cost}}} \wedge \frac{F(V)}{\phi}, \tag{28}$$

where $a(V) = \frac{F(V)}{\phi}$ when the limited liability constraint $F(V) = W(V)$ binds. 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

$$V_0 = \int_0^\infty e^{-\gamma t - \int_0^t \lambda_s ds} W_t dt,$$

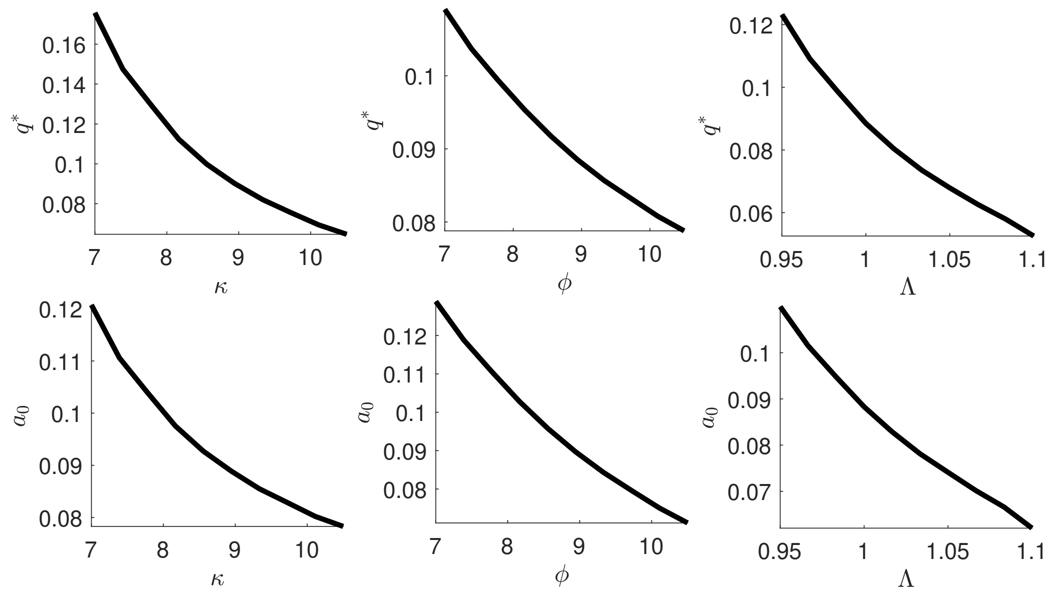

Figure 2: **Comparative Statics.** This figure plots the initial monitoring effort a_0 and screening effort q^* against the parameters ϕ, κ, and Λ.

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 λ_t and so increases the expected time to default; 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 $a^B(q^*)$, 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 (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, $a(V)$ approaches $a^B(q^*)$. As a consequence, the instantaneous default rate λ_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 $\lim_{a_t \to \infty} a_t = a^B$.*

Figure 2 plots optimal screening q^* and initial monitoring a_0 at time $t = 0$ against the cost parameters ϕ and κ and the baseline default intensity Λ. Because monitoring decreases over time so that $a_t < a_0$, the initial level of monitoring proxies the average level of monitoring during the loans' lifetime, in that (ceteris paribus) lower initial monitoring at origination at $t = 0$ implies lower monitoring after origination. The left and center panels of Figure 2 illustrate that monitoring effort a_0 and screening effort q decrease with both the physical costs of monitoring and screening, ϕ and κ. 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 κ 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 ϕ increases, it becomes optimal to curb contracted monitoring and monitoring incentives, leading to lower screening (incentives).

Figure 2 also illustrates that a decrease in the quality of the borrower (or in the quality of the loan), as reflected by the higher baseline default intensity Λ, 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 Λ) borrowers with high priority loans. By contrast, private equity and online lenders finance lower quality (high Λ) 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.[9]

3.2 Implementation

This section shows that the optimal contract can be implemented through time-varying retention of the loan pool by the bank. At origination (i.e., at time $t = 0$), the bank (the agent) retains a fraction β_0 of the loan pool and sells a fraction $1 - \beta_0$ to competitive outside investors. After

[9]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 4 that a similar result obtains in the presence of a credit rating.

origination at times $t \geq 0$, the bank smoothly sells off its stake β_t so that the fraction of the loan pool retained decreases over time. That is, the agent owns a fraction β_t of the overall pool at time t, where β_t is adjusted to provide appropriate incentives W_t.

A per-unit claim on the pool of loans pays the loan rate μ up to default at time τ and therefore has a competitive price

$$D_t = \mu \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du} ds, \tag{29}$$

at any time $t \geq 0$. D_t is linked to credit risk via the instantaneous default intensities λ_s.

Over a short period of time $[t, t+dt]$, the agent receives $\beta_t \mu dt$ in interest payments from the loans. In addition, she sells loans 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:

$$\mu\beta_t - \dot{\beta}_t D_t = c_t. \tag{30}$$

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

$$c_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t > 0. \tag{31}$$

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

$$\mu\beta_t - \dot{\beta}_t D_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t, \tag{32}$$

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 loans 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 β_t and monitoring incentives decrease, which increases default risk and decreases the (pet unit) value of the loan pool D_t.

The following proposition summarizes our results:

Proposition 3 (Implementation)**.** *The optimal contract can be implemented as follows. The agent retains a fraction β_t of the originated loans at time t, whereby a unit stake pays out flow payoff of μ dollars until liquidation at time τ and has 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

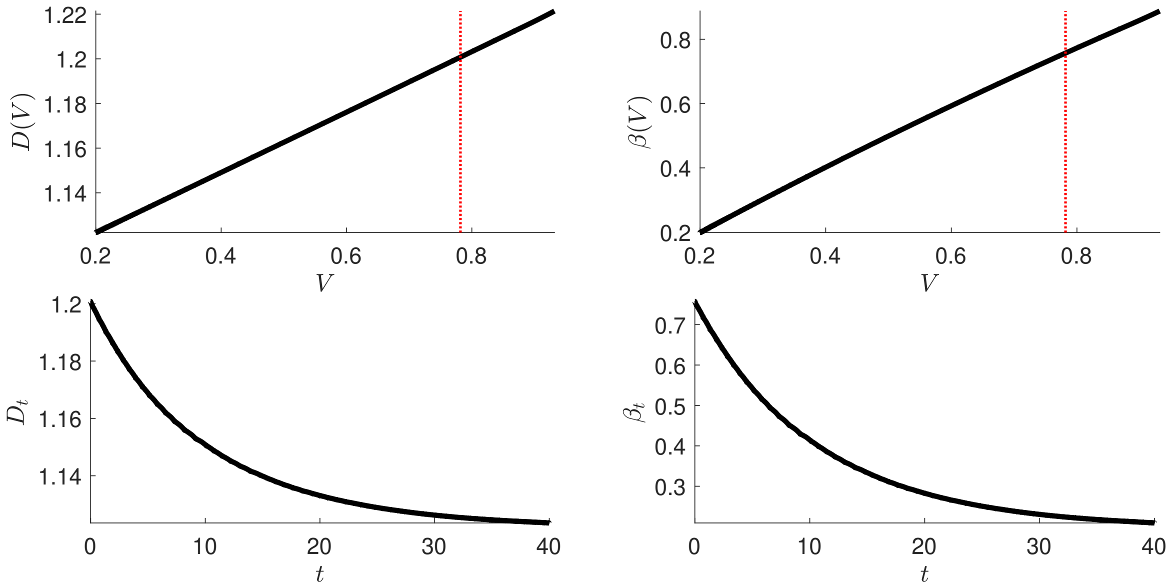

Figure 3: **Implementation of the optimal contract.** The dotted red line depicts the $V = V_0$.

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 $a^B(q) = W^B(q)/\phi$ 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 stake in the loan pool.

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 fraction $\beta^B \in [0,1]$ of the loan pool.*

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 τ^0. At time τ^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 pool of loans until time τ^0. At that time, the agent sells the entire pool to competitive outside investors. This implementation maximizes the agent's exposure to loan performance before time τ^0 (while respecting the principal's limited liability) and allows the agent to capitalize on total surplus $F_{\tau^0} = F(0)$ at time τ^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 τ^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 $dC_t = \mu dt$ 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 full pool of loans until time $\tilde{\tau}$. At time $\tilde{\tau}$, the agent sells the entire pool 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.

4 The effects of credit ratings

Credit rating agencies and their ratings play an important role in alleviating moral hazard between debt investors and debt originators. As we show next, credit ratings 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 rating agencies on credit risk and total surplus, we consider a setting in which the loan pool is rated once at origination, i.e., at time $t = 0$.[10] 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 our setting, the credit rating reveals 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.[11] 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 4 illustrates the effects of credit ratings on the quantities of interest 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 less screening incentives through deferred payouts and therefore holds a lower stake in the pool of loans, 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 the loans 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 4 (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 loan pool (and eliminate front-loading). In addition, and as shown in Proposition 1, the credit rating affects

[10] 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.

[11] 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 the 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.

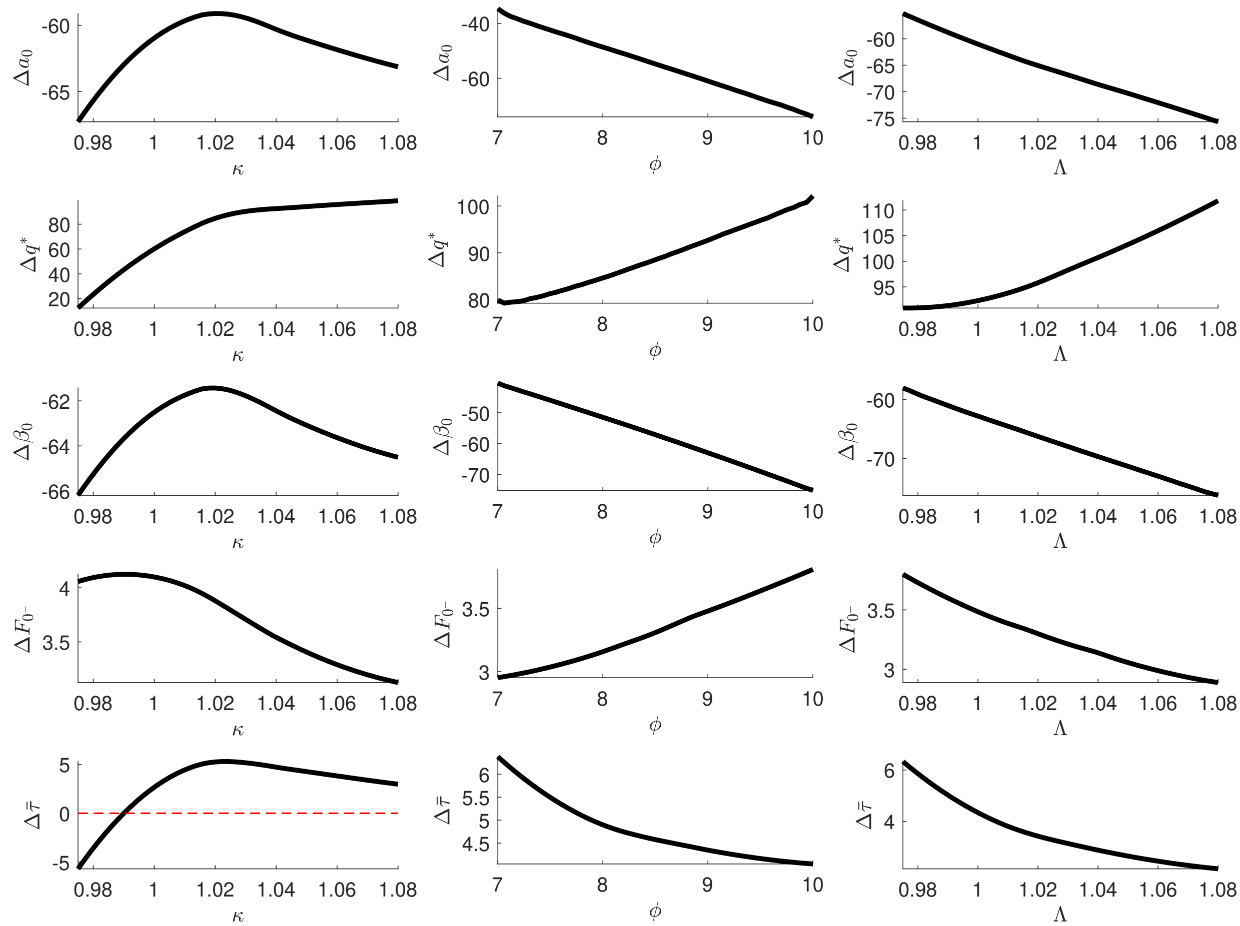

Figure 4: **The effect of credit ratings on effort levels, retention, default risk, and total surplus.** Δy denotes the percentage change in the initial value of the outcome variable y caused by a credit rating, where $y \in \{a_0, q^*, \beta_0, F_{0-}, \bar{\tau}\}$. Outcome variables are plotted as functions of the cost of monitoring κ, the cost of screening ϕ, and the raw default intensity Λ.

the lender's optimal retention level and implies that the bank (loan originator) retains a constant stake in the pool of loans. Notably, we have the following result:

Corollary 3. *Credit ratings imply that loan originators retain a constant stake in the pool of loans and exert a constant level of monitoring effort throughout the life of the pool of loans.*

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 4 shows that when the cost κ 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 to default $\bar{\tau}$ (bottom row in Figure 4). As the cost of screening becomes larger (due e.g. to asset intangibility), moral hazard over screening becomes more important and credit ratings lead to a decrease in default risk.

The fourth row of Figure 4 plots the increase in total surplus ΔF_{0^-} due to a credit rating at origination in percentage terms. A credit rating generates relatively little value when the cost of screening κ is large. The intuition is that when the cost of screening κ is (prohibitively) large, the optimal choice of q is small regardless of whether screening is subject to moral hazard. A low benchmark level of q limits the agency costs related to screening, which, in turn, implies low benefits of removing these agency conflicts by means of a credit rating. Differently, the value of a credit rating is high when the cost ϕ of monitoring is high as it is then optimal to exert relatively more screening and relatively less monitoring. Finally, the value of a credit rating is for high quality loans that are characterized by a low value of exogenous default risk Λ.

5 Loan Maturity and Credit Securitization

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. To compare loans of different maturities, we consider in this section a setting in which loans are securitized at issuance and repeatedly refinanced at maturity with new loans of an identical maturity. Each loan is offered to a borrower seeking long-term financing with terms specifying maturity and future refinancing. We follow Chen, Xu, and Yang (2021) and consider that loans (and therefore the pool of loans) randomly mature with Poisson intensity $\delta > 0$, and all loans mature at the same time. That is, ignoring default, the expected loan maturity is $1/\delta$. The baseline setting corresponds to the case $\delta = 0$, in which loans have infinite maturity. Up to its maturity date, the pool of loans pays coupons at rate μ. When loans mature, firms pay back the face value by issuing new debt with identical face value, coupon, and maturity. That is, firms roll over maturing debt by issuing identical debt claims. Debt is issued at par and firms raise exactly the amount of debt necessary to pay back the face value.[12] We denote by τ^n the n-th maturity date for $n \geq 0$, with $\tau^0 \equiv 0$.

With finite maturity debt, securitization works as follows. The bank securitizes the n-th loans it extends to the firm by offering a contract $\mathcal{C}^n$ to outside investors. When debt matures at time τ^n, the existing contract $\mathcal{C}^n$ between outside investors and the bank ends and the bank extends the $n+1$-th loan to the firm. The bank securitizes these loans by signing a contract $\mathcal{C}^{n+1}$ with outside

[12] This implies that the face value, denoted $\overline{D}$, does not enter the bank's payoff. At the maturity date, the bank is paid back $\overline{D}$ dollars from the firms and, at the same time, lends and transfers $\overline{D}$ dollars to the firms, so that the bank's net payoffs at the maturity date is zero. In other words, the repayment of the face value from the firm and new lending to the firms exactly cancel out.

investors. The contract $\mathcal{C}^n = \{dC_t^n, \hat{a}_t, \hat{q}\}$ represents a claim on the pool of loans originated by the bank and stipulates a profit-sharing rule C^n for the overall loan payments μdt, so that the bank receives dC_t and the investors receive $\mu dt - dC_t$ dollars for times $t \in [\tau^{n-1}, \tau^n]$. In addition, the contract $\mathcal{C}^n$ stipulates a screening level $\hat{q}$ and monitoring levels $\hat{a}_t$ up to the next maturity date (that is, for $t \in [\tau^{n-1}, \tau^n]$). Screening is chosen once at origination at $t = 0$, and the proposed screening level $\hat{q}$ must be the same for all contracts $\mathcal{C}^n$. As in the baseline model with infinite maturity debt, we study incentive compatible contracts that induce $a_t = \hat{a}_t$ and $q = \hat{q}$.

At time τ^{n-1}, the bank can fully commit to any contract $\mathcal{C}^n$ written at that time, but it cannot commit to the contracts $\mathcal{C}^k$ with $k > n$ that it will write in the future. In other words, the bank has full commitment within each of the contracts $\mathcal{C}^n$ but not across contracts. Another interpretation is that at any maturity date τ^n, the bank and outside investors renegotiate contract terms, while there is no renegotiation at any other date.[13]

Note that the first contract $\mathcal{C}^1$ is written before screening q is chosen, but any contract $\mathcal{C}^n$ for $n > 1$ is written after screening effort has been chosen. As a result, when the bank contracts with outside investors at time τ^{n-1} for $n > 1$, there is asymmetric information: the bank knows the actual level of screening q, but outside investors do not.[14] At time τ^{n-1}, outside investors form a belief about the true value of q, denoted q^{n-1}. We consider that this belief coincides with the contracted screening level $\hat{q}$ for any offered contract $\mathcal{C}^n$, in that $\mathbb{P}(q^{n-1} = \hat{q}) = 1$ for $n > 1$.[15] Since we focus on incentive compatible contracts that indeed induce $q = \hat{q}$, these beliefs turn out to be consistent and true.

To characterize the contracting problem with finite debt maturity, we start with the bank's optimization problem at any maturity date τ^n, with $n \geq 1$. At time τ^n, the bank designs a contract $\mathcal{C}^{n+1}$ to maximize total continuation surplus. That is, the bank solves

$$\max_{(a_s)_{s\in[\tau^n,\tau^{n+1}]}} F_{\tau^n} = \max_{(a_s)_{s\in[\tau^n,\tau^{n+1}]}} \int_{\tau^n}^{\infty} e^{-(r+\delta)(s-t)-\int_t^s \lambda_u du} \left(\mu - \frac{\phi a_s^2}{2} - (\gamma - r)W_s + \delta F_{\tau^{n+1}}\right) ds, \tag{33}$$

given the choice of screening, $\hat{q} = q$, and subject to $a_s = \frac{W_s}{\phi}$ (monitoring incentive compatibility). Two observations are in order. First, because—given $\hat{q}$—the contract $\mathcal{C}^{n+1}$ does not affect outside

[13] Note that the long-term contract from the baseline is not renegotiation-proof. Just after screening is chosen, the bank and outside investors could profitably renegotiate the contract terms to implement the "benchmark effort level" $a^B(q)$. However, such renegotiation undermines screening incentives.

[14] Crucially, this information asymmetry is not there when $n = 1$. The first contract $\mathcal{C}^1$ is written at time $t = 0^-$ before screening effort is chosen at time $t = 0$.

[15] Also recall that the contracted screening level $\hat{q}$ must be the same for all offered contracts $\mathcal{C}^n \geq 0$. Thus, at time τ^n for $n \geq 1$, the principal cannot deviate by raising contracted screening effort beyond its level stipulated in contract $\mathcal{C}^0$.

investors' beliefs about the chosen screening level q, the bank has no signalling considerations when designing the contract $\mathcal{C}^{n+1}$. Second, when designing a contract $\mathcal{C}^{n+1}$, the bank treats screening as given (i.e., fixed) and sunk. Therefore, the bank ignores any effects on screening incentives. In turn, treating $q = \hat{q}$ as fixed, the bank solves at τ^n the optimization problem in (19) subject to the monitoring incentive compatibility constraint. As a consequence, any contract $\mathcal{C}^n$ for $n > 1$ merely provides optimal monitoring incentives and screening incentives are determined as the by-product of monitoring incentives. Given $\hat{q} = q$, any contract $\mathcal{C}^n$ for $n > 1$ is stationary and stipulates $V = V^B(q)$, $a = a^B(q)$, and $W = W^B(q)$, leading to $V_t = V^B(q)$ for $t \geq \tau^1$. Thus, the contracts $\mathcal{C}^n$ with $n > 1$ are identical, leading to total surplus $F^B(q)$ from (19) after the first maturity date. Hence, when $V = V^B(q)$, loan maturity δ is no longer payoff relevant and payoffs and incentives are as described in Section 2.2.1.

Next, we turn to the bank's optimization problem at origination (i.e., at time $t = 0^-$). The key difference with any contract $\mathcal{C}^n$ for $n > 1$ is that when designing the contract $\mathcal{C}^1$, the bank takes into account screening incentives, which pin down the actual level of screening through $V_0 = \kappa q$. Thus at origination, the bank solves the optimization in (33) for $n = 0$ subject to the incentive constraints i) $V_0 = \kappa q$ and ii) $a_s = W_s/\phi$. Anticipating the contract structure after the first maturity date, the agent's screening incentives at time $t = 0$ read

$$V_0 = \int_0^\infty e^{-(\gamma+\delta)t - \int_0^t \lambda_s ds}(W_t + \delta V^B(q))dt. \tag{34}$$

At the first maturity date, screening incentives jump to $V^B(q)$ and remain constant thereafter. This is also reflected in the law of motion of V_t which becomes

$$\dot{V}_t = (\gamma + \delta + \lambda_t)V_t - W_t - \delta V^B(q). \tag{35}$$

In contrast, loan maturity has no direct effect on the agent's 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 first maturity date becomes

$$\begin{aligned}(r+\delta)F(V) = \max_{a,W}\Big\{&\mu + \delta F^B(q) - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) \\ &+ F'(V)\big((\gamma+\delta+\lambda)V - W - \delta V^B(q)\big)\Big\},\end{aligned} \tag{36}$$

where $F^B(q)$ is characterized in (19) and the boundary condition $\lim_{V \to V^B(q)} F(V) = F^B(q)$ applies. There are two differences compared with the HJB equation (25). First, the first loan matures with

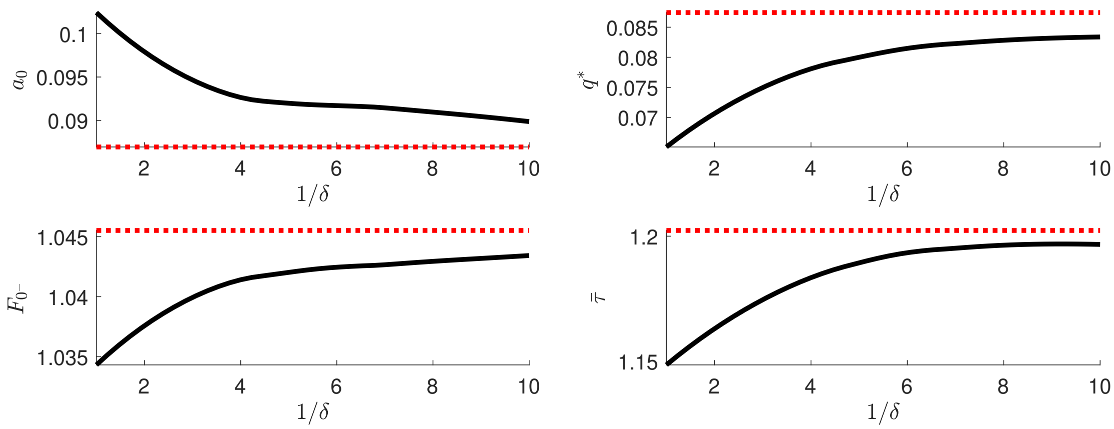

Figure 5: **The effects of debt maturity on screening, monitoring, credit risk, and total surplus.** The dotted red line depicts the outcomes with infinite debt maturity.

intensity δ and total surplus changes by $F^B(q) - F(V)$ when the loan matures. Second, loan maturity affects screening incentives V, so that δ shows up in the flow term that multiplies $F'(V)$.

As in the baseline model, $V_0 > V^B(q)$, and optimal screening effort q^* maximizes total initial surplus $F_{0^-} = F(V_0) - \frac{\kappa q^2}{2}$ subject to the incentive constraint $V_0 = \kappa q$. According to (34), screening incentives V_0 decrease with δ (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 is detrimental to surplus, so 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 (which serves as a proxy for average monitoring), screening effort q^*, and the expected time to default $\bar{\tau}$ (which is inversely related to credit risk) against average debt maturity $1/\delta$. In this figure, firms implement long-term financing by rolling over finite-maturity debt whenever $\delta \neq 0$. Importantly, short maturity undermines the bank's commitment to high powered screening incentives and therefore screening incentives as such. Therefore, as $1/\delta$ decreases, optimal screening incentives and effort q^* decrease. To mitigate this inefficient dilution of screening incentives, the bank receives more front-loaded incentives, in that the bank's initial retention level β_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., $\bar{\tau}$ 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

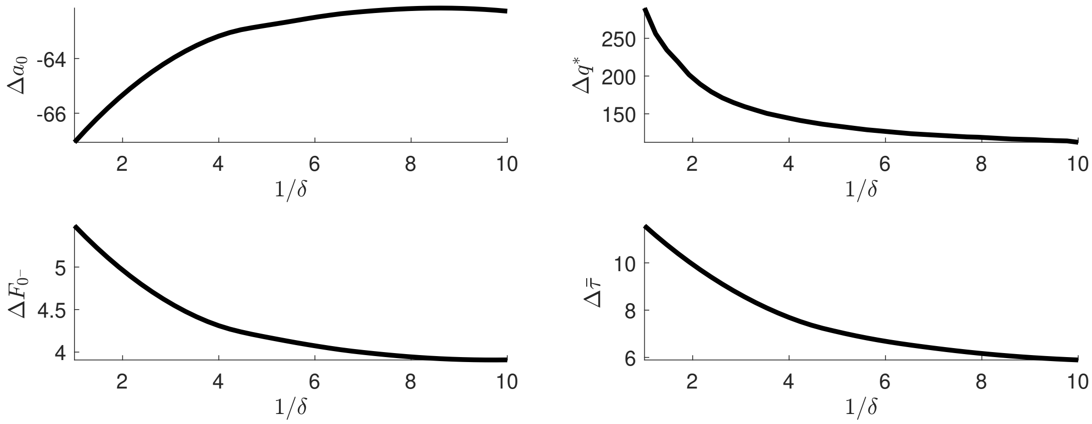

Figure 6: **Debt maturity and the effects of credit ratings.** Δy denotes the percentage change in the initial value of the outcome variable y caused by a credit rating, where $y \in \{a_0, q^*, \beta_0, F_{0-}, \overline{\tau}\}$.

costs.

Section 4 shows that credit ratings can generate value by strengthening screening incentives at the cost of weaker monitoring incentives. Figure 6 displays the value generated by credit ratings at various levels of maturities. As short maturity undermines screening incentives, short-term debt is subject to more severe moral hazard over screening, which has two major consequences. First, a credit rating boosts screening incentives especially for short maturity loans, so that Δq^* decreases with loan maturity $1/\delta$. Second, the value generated by a credit rating, ΔF_{0-}, decreases with debt maturity $1/\delta$, which implies that a credit rating is particularly valuable for firms that implement long-term financing by rolling over short-maturity loans. In addition to its effects on screening, the credit rating reduces (initial) monitoring incentives to a larger extent for short maturity loans. Finally, for all parameters considered, the credit rating reduces credit risk (i.e., $\Delta\overline{\tau}$ is positive). Again the effect is stronger when debt maturity is shorter due to the (positive) effects of credit ratings on screening.

6 Is it optimal to bundle monitoring and screening?

We have so far assumed that the originator of the loans is responsible for both screening and monitoring. In practice, these tasks may be undertaken by separate entities. Some securitized loans are serviced by a third-party serving company and, depending on the specific arraignments, 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 how bundling or separating of 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_t^m and the screener's continuation payoff by W_t^s. 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 5 provides the detailed solution.

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

$$a_t = \frac{W_t^m}{\phi}.$$

Screening effort is determined by the screener's incentive condition

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

where $V_t = \frac{\partial}{\partial q} W_t^s$. Similarly, as in the baseline model and equation (35), we have that

$$\dot{V}_t = (\gamma + \lambda + \delta)V_t - W_t^s - \delta V^B(q).$$

That is,

$$V_0 = \int_0^\infty e^{-(\gamma+\delta)t - \int_0^t \lambda_s ds}(W_t + \delta V^B(q))dt,$$

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 loans mature (with intensity δ), 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, so that $\dot{V}_t < 0$.

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 solve the HJB equation

$$\begin{aligned}(r + \delta)F(V) = \max_{a, W^m, W^s} \Big\{ \mu - \frac{\phi a^2}{2} - (\gamma - r)(W^m + W^s) - \lambda F(V) \\ + \delta F^B(q) + F'(V)\big((\gamma + \lambda + \delta)V - W^s\big)\Big\}, \qquad (37)\end{aligned}$$

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 $W^s = a\phi$. As we show, 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 (37) with respect to W^s yields

$$W^s = F(V) - W^m.$$

Inserting this expression into (37) and simplifying leads to the ordinary differential equation

$$(\gamma + \delta)F(V) = \max_{a, W^m} \left\{ \mu - \frac{\phi a^2}{2} - \lambda F(V) + \delta F(0) + F'(V)\big((\gamma + \lambda + \delta)V - F(V) + W^m\big) \right\}, \quad (38)$$

which is solved subject to $a = W^m/\phi$. As V approaches, the boundary condition $\lim_{V\to 0} F'(V) = -(\gamma - r)$ applies. Due to (38) and $a = W^m/\phi$, the condition $\lim_{V\to 0} F(V) = -(\gamma - r)$ implies

$$\lim_{V\to 0} F(V) = F^B(q) = \max_{a\in[0,\bar{a}]} \left(\frac{\mu - \frac{\phi a^2}{2} - \phi a(\gamma - r)}{\gamma + \lambda} \right), \quad (39)$$

which is expression (19). The maximization in (38) with respect to monitoring effort yields

$$a(V) = \frac{\overbrace{F(V)}^{\substack{\text{Reduction of} \\ \text{default risk}}} \overbrace{-F'(V)V}^{\substack{\text{Screening} \\ \text{incentives}(>0)}} \overbrace{+F'(V)\phi}^{\substack{\text{Screening} \\ \text{disincentives}(<0)}}}{\underbrace{\phi}_{\text{Physical cost}}} \quad \wedge \quad \frac{F(V)}{\phi}. \quad (40)$$

Note that when the limited liability constraint for the screener binds and W^s, then $W^m = F(V)$ and therefore $a(V) = \frac{F(V)}{\phi}$. 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 $V_0 = \kappa q$. We summarize our findings in the following proposition.

Proposition 5 (Separated tasks)**.** *Suppose that monitoring and screening are undertaken by two different agents who are otherwise identical (called the monitor and screener). Denote the monitor's continuation value by W_t^m, the screener's continuation value by W_t^m, and the screener's screening incentives by V_t. In such environments:*

1. *The value function $F(V)$ solves the HJB equation (38) 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 (40).*

3. Optimal screening q^* solves

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

Monitoring and screening are linked by positive and negative terms, denoted as screening incentives and disincentives in equation (40). 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 is given in (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 κ, ϕ, Λ, and $1/\delta$. Under all parameters considered, bundling increases (initial) monitoring (i.e., $\Delta a_0 > 0$), screening ($\Delta q^* > 0$), and total surplus ($\Delta F_{0^-} > 0$).

7 Predictions

Our paper provides several new empirical predictions related to screening and monitoring in debt markets. In the following, we summarize our main predictions.

Prediction 1: *Screening and monitoring are complements so that high screening should be implemented together with high monitoring. Optimal screening and monitoring should decrease with intrinsic borrower quality.*

This first prediction highlights the 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, 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. The model also predicts that a decrease in intrinsic credit quality reduces the marginal benefits, and hence the optimal levels, of screen-

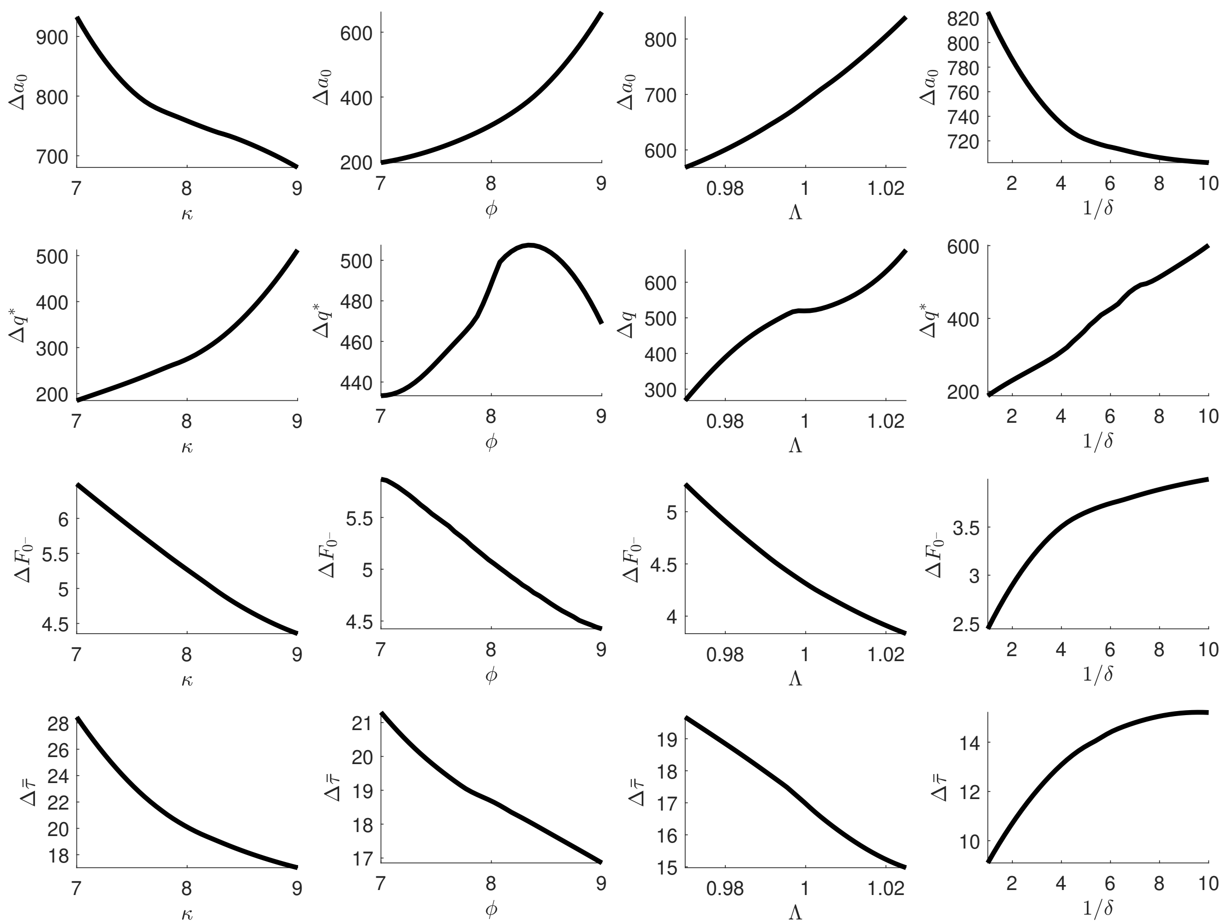

Figure 7: **The effects of bundling screening and monitoring.** Δa_0 denotes the percentage change in monitoring effort at $t = 0$ caused by bundling. Δq^* denotes the percentage change in screening effort q caused by bundling. ΔF_{0^-} denotes the percentage change in total surplus at $t = 0^-$ caused by bundling. $\Delta \overline{\tau}$ denotes the percentage change in the expected time to default caused by bundling.

ing and monitoring, which, in turn, exacerbates credit risk. Our model therefore rationalizes the segmentation observed in credit markets. Indeed, according to our analysis, banks that exert high screening and high monitoring (e.g., via covenants) should finance high quality borrowers with high priority loans. Lenders that implement lower screening (e.g., online lenders) should also implement lower monitoring and finance riskier borrowers. Our model also rationalizes 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.

Prediction 2: *When securitized loans are rated, the originator's retention level in the pool of loans changes less (and in particular decreases less) over time than when securitized loans are not rated.*

This second prediction follows from Proposition 3 and Corollary 2 and relates to the optimal size of the stake retained by the loan originator in order to maximize total surplus (and minimize agency frictions). According to the model, the synergies between monitoring and screening 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 bank's incentives by means of delayed payouts are especially strong at origination and decrease over time. Accordingly, monitoring incentives decrease over time. Credit ratings effectively remove moral hazard over screening. As a result, originators only need to be incentivized to exert monitoring effort, which can be done with a constant stake in the pool of securitized loans.

Prediction 3: *Credit ratings are particularly valuable for high quality borrowers, for firms that implement long-term financing by rolling over short-maturity debt, and for assets or borrowers for which screening is not costly but monitoring is costly. Credit ratings may lead to an increase in credit risk due to their adverse effects on monitoring incentives.*

The third prediction follows from the fact that credit ratings remove moral hazard over screening so that the loan originator does not need to be incentivized to screen borrowers. The loan originator therefore optimally holds a lower stake in the pool of loans. This, in turn, leads to lower monitoring incentives (and levels) and, potentially, to higher credit risk notably when the cost of screening is low. When monitoring is costly, it is optimal to implement low levels of monitoring so that credit ratings have little effect on actual levels of monitoring and on credit risk. Lastly, when the intrinsic quality of the borrower is low, it is optimal to implement low monitoring and screening, which reduces the value of credit ratings.

Prediction 4: *For firms that implement long-term financing by rolling over short-maturity debt,* *an increase in debt maturity reduces credit risk and the value of credit ratings.*

The fourth prediction follows from the fact that short maturity undermines the bank's commitment to high powered screening incentives and therefore screening incentives as such. Therefore, as debt maturity decreases, optimal screening incentives and effort decrease, thereby increasing credit risk and rendering credit ratings more valuable. Therefore we expect the incidence of observing ratings to decrease with debt maturity.

Prediction 5: *Default risk should be lower when loan originators are responsible for both screening and monitoring. The effects of bundling on default risk should increase with debt maturity and decrease with credit quality.*

The fifth prediction follows from the fact that bundling monitoring with screening increases the skin in the game of the loan originator, leading to higher screening and monitoring efforts. As a result, of higher effort, default risk decreases, in line with the evidence in Demiroglu and James (2012). The additional predictions are novel.

8 Conclusion

We study a dynamic moral hazard problem in which a bank originates a pool of loans to sell them to investors. The bank controls the loans' 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 agent's payoff to loan performance. As screening occurs only once at the origination of the loans, the agent's incentives are front-loaded and stronger shortly after origination. The optimal contract can be implemented via time-decreasing retention of a stake within the loan pool 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 in credit securitization 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 credit ratings are particularly valuable for high quality borrowers and for firms that implement long-term financing by rolling over short-maturity debt. By removing moral hazard over screening, credit ratings reduce the size of the stake that the loan originator should retain in the pool of securitized loans and imply that this stake should be constant until the loans mature. A lower stake in the securitized loans reduces monitoring incentives so that credit ratings may lead to an increase in credit risk.

References

Biais, B., T. Mariotti, G. Plantin, and J.-C. Rochet (2007). Dynamic security design: Convergence to continuous time and asset pricing implications. *Review of Economic Studies 74*(2), 345–390.

Biais, B., T. Mariotti, J.-C. Rochet, and S. Villeneuve (2010). Large risks, limited liability, and dynamic moral hazard. *Econometrica 78*(1), 73–118.

Celik, S., G. Demirtaş, and M. Isaksson (2019). Corporate bond markets in a time of unconventional monetary policy. *OECD Capital Market Series*.

Chen, H., Y. Xu, and J. Yang (2021). Systematic risk, debt maturity, and the term structure of credit spreads. *Journal of Financial Economics 139*, 770–799.

Cordell, L., M. Roberts, and M. Schwert (2021). Clo performance. *Working Paper, University of Pennsylvania*.

Daley, B., B. Green, and V. Vanasco (2020). Securitization, ratings, and credit supply. *Journal of Finance 75*(2), 1037–1082.

DeMarzo, P. and D. Duffie (1999). A liquidity-based model of security design. *Econometrica 67*(1), 65–99.

DeMarzo, P. and Z. He (2021). Leverage dynamics without commitment. *The Journal of Finance 76*(3), 1195–1250.

DeMarzo, P. M., M. J. Fishman, Z. He, and N. Wang (2012). Dynamic agency and the q theory of investment. *The Journal of Finance 67*(6), 2295–2340.

DeMarzo, P. M. and Y. Sannikov (2006). Optimal security design and dynamic capital structure in a continuous-time agency model. *Journal of Finance 61*(6), 2681–2724.

Demiroglu, C. and C. James (2012). How important is having skin in the game? originator-sponsor affiliation and losses on mortgage-backed securities. *Review of Financial Studies 25*(11), 3217–3258.

Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *Review of Economic Studies 51*(3), 393–414.

Drucker, S. and M. Puri (2009). On loan sales, loan contracting, and lending relationships. *Review of Financial Studies 22*(7), 2835–2872.

Feng, F. Z., C. R. Taylor, M. M. Westerfield, and F. Zhang (2021). Setbacks, shutdowns, and overruns. *Available at SSRN 3775340*.

Feng, F. Z. and M. M. Westerfield (2021). Dynamic resource allocation with hidden volatility. *Journal of Financial Economics 140*(2), 560–581.

Gorton, G. B. and G. G. Pennacchi (1995). Banks and loan sales marketing nonmarketable assets. *Journal of Monetary Economics 35*(3), 389–411.

Green, B. and C. R. Taylor (2016). Breakthroughs, deadlines, and self-reported progress: contracting for multistage projects. *American Economic Review 106*(12), 3660–99.

Gryglewicz, S. and S. Mayer (2021). Delegated monitoring and contracting. *Working Paper*.

Gryglewicz, S., S. Mayer, and E. Morellec (2020). Agency conflicts and short-versus long-termism in corporate policies. *Journal of Financial Economics 136*(3), 718–742.

Halac, M. and A. Prat (2016, October). Managerial attention and worker performance. *American Economic Review 106*(10), 3104–32.

Hartman-Glaser, B., T. Piskorski, and A. Tchistyi (2012). Optimal securitization with moral hazard. *Journal of Financial Economics 104*(1), 186–202.

He, Z. (2009). Optimal executive compensation when firm size follows geometric brownian motion. *The Review of Financial Studies 22*(2), 859–892.

He, Z. (2011). A model of dynamic compensation and capital structure. *Journal of Financial Economics 100*(2), 351–366.

He, Z. (2012). Dynamic compensation contracts with private savings. *Review of Financial Studies 25*(5), 1494–1549.

Hoffmann, F., R. Inderst, and M. M. Opp (2020). The economics of deferral and clawback requirements. *Working Paper*.

Hoffmann, F., R. Inderst, and M. M. Opp (2021). Only time will tell: A theory of deferred compensation. *Review of Economic Studies 88*(3), 1253–1278.

Holmstrom, B. (1989). Agency costs and innovation. *Journal of Economic Behavior & Organization 12*(3), 305–327.

Hu, Y. and F. Varas (2021). Intermediary financing without commitment.

Ivashina, V. and B. Vallée (2021). Weak credit covenants. *Working Paper, Harvard Business School*.

Kundu, S. (2021). The anatomy of collateralized loan obligations: On the origins of covenants and contract design. Technical report.

Malamud, S., H. Rui, and A. Whinston (2013). Optimal incentives and securitization of defaultable assets. *Journal of Financial Economics 107*(1), 111–135.

Malenko, A. (2019). Optimal dynamic capital budgeting. *Review of Economic Studies 86*(4), 1747–1778.

Marinovic, I. and M. Szydlowski (2020). Monitor reputation and transparency. *Available at SSRN 3703870*.

Marinovic, I. and F. Varas (2019). Ceo horizon, optimal pay duration, and the escalation of short-termism. *Journal of Finance 74*(4), 2011–2053.

Mayer, S. (2020). Financing breakthroughs under failure risk. *Working Paper*.

Pagano, M. and P. Volpin (2010). Credit ratings failures and policy options. *Economic Policy 25*, 401–431.

Pennacchi, G. G. (1988). Loan sales and the cost of bank capital. *Journal of Finance 43*, 375–396.

Piskorski, T. and M. M. Westerfield (2016). Optimal dynamic contracts with moral hazard and costly monitoring. *Journal of Economic Theory 166*, 242–281.

Sannikov, Y. (2008). A continuous-time version of the principal-agent problem. *Review of Economic Studies 75*(3), 957–984.

Varas, F. (2018). Managerial short-termism, turnover policy, and the dynamics of incentives. *The Review of Financial Studies 31*(9), 3409–3451.

Varas, F., I. Marinovic, and A. Skrzypacz (2020). Random inspections and periodic reviews: Optimal dynamic monitoring. *The Review of Economic Studies 87*(6), 2893–2937.

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

$$\max_{a_t \in [0,\bar{a}]} \left(a_t W_t - \frac{\phi a_t^2}{2} \right), \tag{41}$$

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 (41). The second order condition to (41), i.e., $\frac{\partial^2}{\partial a_t^2} \left(a_t W_t - \frac{\phi a_t^2}{2} \right) = -\phi < 0$, is satisfied.

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

$$\max_{q \in [0,\bar{q}]} \left(W_0(q) - \frac{\kappa q^2}{2} \right), \tag{42}$$

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 (42) satisfies $q > 0$.

Now observe that

$$q = \min \left\{ \frac{V_0(q)}{\kappa}, \bar{q} \right\} \tag{43}$$

is the unique solution to (42) if

$$\frac{\partial^2}{\partial q^2} \left(W_0(q) - \frac{\kappa q^2}{2} \right) = \frac{\partial}{\partial q} V_0(q) - \kappa < 0 \tag{44}$$

holds for any $q \in [0, \bar{q}]$, in which case the objective in (42) is strictly concave over the entire interval $[0, \bar{q}]$ and the first order approach is valid. When optimal screening effort is interior, condition (43) simplifies to (12), i.e., $q = V_0/\kappa$, which is the first order condition to (42).

In what follows, we provide a sufficient condition for (44) 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 (44) 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

$$Y_t(q) = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} 2V_s(q)ds \tag{45}$$

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

$$V_t(q) = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} W_s(q)ds \tag{46}$$

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

$$\lambda_t = \Lambda - a_t - q \geq \Lambda - \bar{a} - \bar{q}. \tag{47}$$

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

$$W_t < W^{max} = \frac{\mu}{r + \Lambda - \bar{a} - \bar{q}}, \tag{48}$$

whereby $W^{\max}$ is the total surplus evaluated under discount rate r, that is,

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

when monitoring effort is costless (so that $a_s = \bar{a}$).

Using these two relations (47) and (48) as well as (46), we obtain that

$$V_t(q) < \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} W^{max} ds = \frac{W^{max}}{\gamma + \Lambda - \bar{a} - \bar{q}} < \frac{\mu}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})} \tag{49}$$

Using this inequality (49) and the integral representation in (45), we obtain that

$$Y_t(q) < \frac{2\mu}{(\gamma + \Lambda - \bar{a} - \bar{q})^2 (r + \Lambda - \bar{a} - \bar{q})}.$$

As a result, a sufficient condition for (44), i.e., for

$$Y_0(q) < \kappa$$

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

$$\kappa \geq \frac{2\mu}{(\gamma + \Lambda - \bar{a} - \bar{q})^2 (r + \Lambda - \bar{a} - \bar{q})}. \tag{50}$$

That is, when (50) holds, the first order approach is valid and (43) or, equivalenty, (12) (due to $q < \bar{q}$) pins down screening effort. Note that (50) 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

$$F_t = \underbrace{\int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}(\mu ds - dC_s)}_{=P_t} + \underbrace{\int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du}\left(dC_s - \frac{\phi a_s^2}{2}ds\right)}_{=W_t},$$

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

$$P_t = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}(\mu ds - dC_s)$$

is the principal's continuation payoff and

$$W_t = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du}\left(dC_s - \frac{\phi a_s^2}{2}ds\right)$$

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

$$dP_t = (r + \lambda_t)P_t dt - \mu dt + dC_t \tag{51}$$

$$dW_t = (\gamma + \lambda_t)W_t dt + \frac{\phi a_t^2}{2}dt - dC_t \tag{52}$$

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

$$\begin{aligned} dF_t &= dP_t + dW_t && (53) \\ &= (r + \lambda_t)P_t dt - \mu dt + dC_t + (\gamma + \lambda_t)W_t dt - dC_t + \frac{\phi a_t^2}{2}dt \\ &= (r + \lambda_t)(\underbrace{P_t + W_t}_{=F_t})dt - \mu dt + \frac{\phi a_t^2}{2}dt - (\gamma - r)W_t dt. && (54) \end{aligned}$$

We can integrate (53) over time, t, to get

$$F_t = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}\left(\mu - \frac{\phi a_s^2}{2} - (\gamma - r)W_s\right)ds, \tag{55}$$

which is (23) from the main text.

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

$$F_0 - \frac{\kappa q^2}{2}, \tag{56}$$

where F_0 is characterized in (55). 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

$$rF_t = \max_{W_t \in [0,F_t], a_t \geq 0} \left(\mu - \frac{\phi a_t^2}{2} - (\gamma - r)W_t + \dot{F}_t - \lambda_t F_t \right),$$

which is solved subject to the monitoring incentive condition (7) and where $\dot{F}_t = \frac{dF_t}{dt}$. 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 $\dot{F}_t = 0$ and we can omit time sub-scripts (i.e., we write $F_t = F^B(q)$). In turn, the HJB equation simplifies to

$$rF^B(q) = \max_{W \in [0,F^B(q)], a \geq 0} \left(\mu - \frac{\phi a^2}{2} - (\gamma - r)W_t - \lambda F^B(q) \right) \tag{57}$$

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

$$a^B(q) = \frac{F^B(q) - \phi(\gamma - r)}{\phi}, \tag{58}$$

and the optimal 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 (58) 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 (51) and $dW_t = 0$ imply then that payouts to the agent take the form $dC_t = c^B(q)dt$ with

$$c^B(q) = (\gamma + \lambda^B(q))W^B(q) + \frac{\phi(a^B(q)^2}{2}. \tag{59}$$

That is, payouts to the agent are smooth and positive.

The objective (56) can be rewritten as

$$F^B(q) - \frac{\kappa q^2}{2}. \tag{60}$$

At time $t = 0$, the agent chooses screening effort $q \in [0, \bar{q}]$ to maximize (60), 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 (57) holds. Using the envelope theorem, we can differentiate both sides of (57) with respect to $q = q^B$ to obtain under the optimal controls $(W^B(q), a^B(q))$

$$(r + \lambda)\frac{\partial F^B(q)}{\partial q} = F^B(q) \quad \Longleftrightarrow \quad \frac{\partial F^B(q)}{\partial q} = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} > 0 \tag{61}$$

for $q = q^B$. Utilizing above expression for $\frac{\partial F^B(q)}{\partial q}$, the first order condition to (21),which is $\frac{\partial F^B(q)}{\partial q} =$

κq, becomes (22), as desired. Also observe that $\frac{\partial F^B(q)}{\partial q}$ and (58) 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 $\frac{\partial^3}{\partial q^3} F^B(q) > 0$. 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,

$$q^B = \min \left\{ q \in [0, \bar{q}] : \frac{\partial F^B(q)}{\partial q} = \kappa q \right\}. \tag{62}$$

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

C Proof of Proposition 2

The proof is split in four 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 $q \in (0, \bar{q})$.

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 (55):

$$F_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} \left(\mu - \frac{\phi a_s^2}{2} - (\gamma - r) W_s \right) ds.$$

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

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

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 $W \in [0, F(V)]$ 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

$$a(V) = \frac{F(V) - F'(V)[V+\phi] - (\gamma - r)\phi}{\phi} \quad \wedge \quad W(C), \tag{63}$$

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

$$V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}. \tag{64}$$

It follows that when $V = V^B(q)$, the continuation surplus is $F^B(q)$. That is, the ODE (25) is solved subject to

$$\lim_{V\to V^B(q)} F(V) = F^B(q), \tag{65}$$

where the limit accounts for the possibility that the state variable V may never reach $V^B(q)$, even if $\lim_{t\to\infty} V_t = V^B(q)$ which is shown in the next Part (i.e., Part II) of the proof. Also note that optimal effort $a(V)$ satisfies in the limit $\lim_{V\to V^B(q)} a(V) = a^B(q)$.

C.2 Part II

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 $t = 0$

$$\begin{aligned}\mathcal{L} &= \underbrace{\int_0^\infty e^{-rt-\int_0^t \lambda_u du}\left(\mu - (\gamma - r)W_t - \frac{\phi a_t^2}{2}\right)dt}_{=F_0} + \ell\Big(\kappa q - \underbrace{\int_0^\infty e^{-\gamma t - \int_0^t \lambda_u du} W_t dt}_{=V_0}\Big) \\ &= F_0 + \ell(\kappa q - V_0). \end{aligned} \tag{66}$$

where ℓ 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 insert into the objective function.

Next, we rewrite (53) as

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

which can be integrated over time to obtain

$$F_t = \int_t^{\infty} e^{-r(s-t)} \left(\mu - \frac{\phi a_s^2}{2} - (\gamma - r) W_s - \lambda_s F_s \right) ds. \tag{67}$$

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

$$V_t = \int_t^{\infty} e^{-\gamma(s-t)} \left(W_s - \lambda_s V_s \right) ds. \tag{68}$$

Using (67) and (68), we can rewrite the Lagrangian (66) as

$$\mathcal{L} = \int_0^{\infty} e^{-rt} \left(\mu - (\gamma - r) W_t - \frac{\phi a_t^2}{2} - \lambda_t F_t \right) dt + \ell \left(\kappa q - \int_0^{\infty} e^{-\gamma t} (W_t - \lambda_t V_t) dt \right), \tag{69}$$

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:

$$e^{-rt}(F_t - (\gamma - r)\phi - \phi a_t) - \ell e^{-\gamma t}(\phi + V_t) = 0 \tag{70}$$

Multiplying both sides of (70) by e^{rt}, we obtain

$$F_t - (\gamma - r)\phi - \phi a_t - \ell e^{-(\gamma - r)t}(\phi + V_t) = 0. \tag{71}$$

We can solve (71) for

$$a_t = \frac{F_t - (\gamma - r)\phi - \ell e^{-(\gamma - r)t}(V_t + \phi)}{\phi}. \tag{72}$$

Taking the limit $t \to \infty$ in (72) leads to

$$\lim_{t \to \infty} a_t = \lim_{t \to \infty} \left(\frac{F_t - (\gamma - r)\phi}{\phi} \right), \tag{73}$$

as V_t is bounded (see inequality (49) in the proof of Lemma 1 and note that by definition, $V_t \geq 0$).

The expression for F_t (e.g., (23) and (67)) and the relation (73) imply that

$$\lim_{t \to \infty} F_t = \hat{F} \quad \text{and} \quad \lim_{t \to \infty} a_t = \hat{a}$$

for constants $\hat{F}$ and $\hat{a}$. Note that by (73),

$$\hat{a} = \frac{\hat{F} - (\gamma - r)\phi}{\phi} \tag{74}$$

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

$$\hat{F} = \frac{\mu - (\gamma - r)\phi\hat{a} - \frac{\phi\hat{a}^2}{2}}{r + \Lambda - \hat{a} - q}. \tag{75}$$

Now note that (74) and (75) 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

$$\lim_{t\to\infty} V_t = V^B(q) \quad \text{and} \quad \lim_{t\to\infty} \dot{V}_t = 0. \tag{76}$$

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 $\dot{V}_t = 0$, as otherwise (76) would not hold.

That is, when $V_0 = \kappa q > V^B(q)$, it follows that $\dot{V}_t < 0$, with convergence according to (76). Likewise, when $V_0 = \kappa q < V^B(q)$, it follows that $\dot{V}_t > 0$, with convergence according to (76). In the knife-edge case $V_0 = \kappa q = V^B(q)$, it holds that $V_t = V^B(q)$ and $\dot{V}_t = 0$.

Last, we characterize the limit $\lim_{V\to V^B(q)} F'(V)$. Recall that $\lim_{V\to V^B(q)} F(V) = F^B(q)$ and $\lim_{V\to V^B(q)} a(V) = a^B(q)$. We know 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 (63), $\lim_{V\to V^B(q)} a(V) = a^B(q)$ becomes equivalent to

$$\lim_{V\to V^B(q)} F'(V) = 0. \tag{77}$$

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

$$F(V_0) - \frac{\kappa q^2}{2} \quad \text{s.t.} \quad V_0 = \kappa q. \tag{78}$$

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 (78) is

$$\frac{\partial F(V_0)}{\partial q}|_{q=q^*} + F'(V_0)\kappa = \kappa q^* \tag{79}$$

where analogously to (61):

$$\frac{\partial F(V_0)}{\partial q} = \frac{F(V_0)}{r + \Lambda - a(V_0) - q}. \tag{80}$$

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

$$\kappa q^* = V_0 \leq V^B(q^*) = \frac{W^B(q^*)}{\gamma + \Lambda - a^B(q^*) - q^*}, \tag{81}$$

where the last equality follows (64). 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

$$\kappa q^* = V_0 \leq \frac{W^B(q^*)}{\gamma + \Lambda - a^B(q^*) - q^*} < \frac{F^B(q^*)}{r + \Lambda - a^B(q^*) - q^*}, \tag{82}$$

where the first inequality follows (81) 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 $\frac{\partial^3}{\partial q^3} F^B(q) > 0$.[16] Define the function $G(q) = F^B(q) - \frac{\kappa q^2}{2}$, which is the objective function in (21). Using (61), we obtain

$$G'(q) = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q.$$

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 $q \in [0, q^B)$.[17]

Next, consider the continuous function

$$K(q) = V^B(q) - \kappa q \tag{83}$$

Note that $a^B(q)$ and $W^B(q)$ increase with q (see Proposition 1). Thus, by (64), the function $V^B(q)$ is strictly convex, implying that $K(q)$ is strictly convex too. Observe that according to (82)

$$K(q) = V^B(q) - \kappa q = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q} - \kappa q < \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q = G'(q).$$

Because i) $G'(q)$ has a unique root on $[0, q^B]$, ii) because $K(q) < G'(q)$, iii) because $K(q)$ is convex,

[16]To see this, note that according to (61), $F^B(q)$ increases with q and according to Proposition 1, $a^B(q)$ increases with q too.

[17]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 (62).

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 $q \in (\hat{q}, q^B]$.

Next, note that for $q = \bar{q}$:

$$K(\bar{q}) = \frac{W^B(\bar{q})}{\gamma + \Lambda - a^B(\bar{q}) - \bar{q}} = \frac{a^B(\bar{q})\phi}{\gamma + \Lambda - a^B(\bar{q}) - \bar{q}} \leq \frac{\bar{a}\phi}{\gamma + \Lambda - \bar{a} - \bar{q}} \leq 0,$$

where the second equality uses (7) and that the incentive constraint for monitoring effort binds, the first inequality uses $a^B(\bar{q}) \leq \bar{a}$, 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 $G(q)$ 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 $V_t = V^B(\hat{q}) = \kappa\hat{q}$ 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 $q^* = \hat{q}$. Take $\varepsilon > 0$ and set $q^\varepsilon = q^* + \varepsilon$ so that $q^\varepsilon < q^B$. Because of $q^* < q^B$, it follows that

$$\frac{\partial}{\partial q^*}\left(F^B(q^*) - \frac{\kappa(q^*)^2}{2}\right) > 0, \tag{84}$$

where $F^B(q^*) - \frac{\kappa(q^*)^2}{2}$ is total surplus under the optimal choice of q, i.e., $q = q^* = \hat{q}$.

Under the screening level $q^\varepsilon = q^* + \varepsilon$, it follows that $\kappa q^\varepsilon = V_0 > V^B(q^\varepsilon)$. Denote the value function under screening level q^ε by $F(V)$. The total surplus under screening level q^ε is

$$\begin{aligned} F(V_0) - \frac{\kappa(q^\varepsilon)^2}{2} &= F^B(q^\varepsilon) + F'(V^B(q^\varepsilon))\varepsilon + o(\varepsilon^2) - \frac{\kappa(q^\varepsilon)^2}{2} = F^B(q^\varepsilon) + o(\varepsilon^2) - \frac{\kappa(q^\varepsilon)^2}{2}, \\ &= \left(F^B(q^*) - \frac{\kappa(q^*)^2}{2}\right) + \frac{\partial}{\partial q^*}\left(F^B(q^*) - \frac{\kappa(q^*)^2}{2}\right)\varepsilon + o(\varepsilon^2), \end{aligned} \tag{85}$$

which — by (84) — exceeds $F^B(q^*) - \frac{\kappa(q^*)^2}{2}$ for $\varepsilon > 0$ sufficiently small. The second equality uses that given screening level q^ε, $\lim_{V \to V^B(q^\varepsilon)} F'(V) = 0$ (see (77)). 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 (63) implies that $F'(V) < 0$, because $F'(V) \geq 0$ would imply $a(V) < F(V)/\phi$ and $W(V) < F(V)$.

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, let us show that the value function is strictly decreasing and concave for all $V > V^B(q)$. As the principal's limited liability constraint does not bind, optimal effort $a(V)$ solves the first order condition $\frac{\partial F(V)}{\partial a} = 0$. We can then invoke the envelope theorem and totally differentiate the HJB equation (25) under the optimal controls with respect to V which yields

$$F''(V) = \frac{-(\gamma - r)F'(V)}{(\gamma + \lambda)V - W}. \tag{86}$$

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 (86) that either $F'(V), F''(V) < 0$ for all $V > V^B(q)$ or $F'(V), F''(V) \geq 0$ for all $V > V^B(q)$.

If it were $F'(V), F''(V) \geq 0$ for all $V > V^B(q)$, then $F(V) \geq F^B(q)$ for $V \geq V^B(q)$, as $F^B(q) = F(V^B(q))$ by means of (65). 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, it follows that $F'(V), F''(V) < 0$ for all $V > V^B(q)$.

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

$$c_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t. \tag{87}$$

If $F_t = W_t$, note that according to (53), $\dot{F}_t = (\gamma + \lambda_t)F_t - \mu + \frac{\phi a_t^2}{2}$. Inserting the law of motion $\dot{F}_t = \dot{W}_t$ into (87) yields $c_t = \mu > 0$.

Next, consider $V = V_t$ with $W_t < F_t$. Then, $a(V) = \frac{F(V) - F'(V)[V+\phi] - (\gamma - r)\phi}{\phi}$ and $a'(V) = \frac{-F''(V)}{\phi} > 0$, 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 (87) 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 (63), optimal monitoring effort (if interior) satisfies

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

so that

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

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

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

solves the ODE

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

subject to the boundary condition

$$\lim_{V \to V^B(q)} D'(V) = 0 \quad \Longleftrightarrow \quad \lim_{V \to V^B(q)} D(V) = \frac{\mu}{r + \Lambda - a^B(q) - q}.$$

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

$$\mu\beta(V) - \beta'(V)\dot{V}D(V) = (\gamma + \Lambda - a(V) - q)W(V) + \frac{\phi a_t^2}{2} - W'(V)\dot{V}, \tag{88}$$

which is solved subject to

$$\lim_{V \to V^B(q)} \beta'(V) = 0 \quad \Longleftrightarrow \quad \lim_{V \to V^B(q)} \beta(V) = \frac{c^B(q)}{\mu}.$$

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 (88), as desired.

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 $\beta^B(q) = c^B(q)/\mu$ 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

$$rF(V) = \max_{W \in [0,F(V)], a \in [0,\bar{a}]} \left\{ \mu - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)((\gamma + \lambda)V - W) \right\}. \tag{89}$$

The maximization with respect to monitoring effort, a, yields that, if interior, optimal monitoring effort is

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

The maximization with respect to the agent's deferred compensation yields that

$$W(V) \begin{cases} = 0 & \text{if } F'(V) > -(\gamma - r) \\ \in [0, F(V)] & \text{if } F'(V) = -(\gamma - r) \\ = F(V) & \text{if } F'(V) < -(\gamma - r). \end{cases} \tag{90}$$

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 $\dot{V}_t < 0$ at all times $t \geq 0$, in that

$$\dot{V} = (\gamma + \lambda)V - W(V) < 0.$$

Owing to (90), this requires that $W(V) > 0$ for $V > 0$ and therefore $F'(V) \leq -(\gamma - r)$ for $V > 0$, owing to (90).

Thus, it is (at least) weakly optimal to stipulate $W(V) = F(V)$, which we can insert into the HJB equation (89) to obtain

$$\gamma F(V) = \max_{a \in [0,\bar{a}]} \left\{ \mu - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)((\gamma + \lambda)V - F(V)) \right\}. \tag{91}$$

Using the envelope theorem, we totally differentiate the HJB equation (91) (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 $\dot{V} = (\gamma + \lambda)V - F(V) < 0$, $F''(V) < 0$. That is, $F(V)$ is strictly concave for $V > 0$. If there exists now $\hat{V} > 0$ with $F'(\hat{V}) = -(\gamma - r)$, then there exists $0 < V' < \hat{V}$ 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 (90) that $F'(0) \geq -(\gamma - r)$. As $F'(V) < -(\gamma - r)$ 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 μ.

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

$$dW = (\gamma + \lambda)Wdt + \frac{\phi a^2}{2}dt - dC,$$

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 (53) for $W(V) = F(V)$:

$$dW = (\gamma + \lambda)Wdt + \frac{\phi a^2}{2}dt - dC = (\gamma + \lambda)Fdt + \frac{\phi a(V)^2}{2}dt - \mu dt = dF_t,$$

yielding

$$dC = \mu dt,$$

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_t^s and payouts to the monitor as dC_t^m. The contracts to screener and monitor stipulate payouts $\{dC_t^s\}$ and $\{dC_t^m\}$ 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

$$W_t^s = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} dC_s^s$$

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

$$W_t^m = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} \left(dC_s^m - \frac{\phi a_s^2}{2} ds \right),$$

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

$$\begin{aligned} dW_t^s &= (\gamma + \lambda_t) W_t^s dt - dC_t^s \\ dW_t^m &= (\gamma + \lambda_t) W_t^m dt - dC_t^m + \frac{\phi a_t^2}{2} dt \end{aligned}$$

As dC_t^s and dC_t^m are not sign-restricted, we can treat W_t^s and W_t^m as control variables in the dynamic optimization problem, while dropping the controls dC_t^s and dC_t^m.

Analogous to the baseline, we define the screener's screening incentives at time t as

$$V_t = \int_t^\infty e^{-(\gamma+\delta)(s-t)-\int_t^s \lambda_u du} (W_s^s + \delta V^B(q)) ds, \tag{92}$$

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 $V_t = W_t^s = 0$ at all times $t \geq 0$.

Thus,

$$dV_t = (\gamma + \lambda_t + \delta) V_t dt - W_t^s dt. \tag{93}$$

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 is pinned down by

$$a_t = \frac{W_t^m}{\phi},$$

which is analogous to (7).

The optimal contracts to both the screener and monitor are designed to dynamically maximize total surplus

$$F_t = \underbrace{\int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du} (\mu ds - dC_s^s - dC_s^m)}_{=P_t} + W_t^m + W_t^s.$$

Total surplus F_t can be rewritten (using arguments analogous to the ones that lead to (55)) as

$$F_t = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du} \left(\mu - \frac{\phi a_s^2}{2} - (\gamma - r)(W_s^s + W_s^m) \right) ds.$$

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

$$(r+\delta)F(V) = \max_{a,W^m,W^s} \Big\{\mu - \frac{\phi a^2}{2} - (\gamma - r)(W^m + W^s) - \lambda F(V) \tag{94}$$

$$+ \delta F(0) + F'(V)\big((\gamma+\lambda+\delta)V - W^s\big)\Big\}, \tag{95}$$

which is (37). 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 $W^s = a\phi$.

The maximization with respect to the screener's deferred compensation W^s yields that

$$W^s(V) \begin{cases} = 0 & \text{if } F'(V) > -(\gamma - r) \\ \in [0, F(V) - W^m(V)] & \text{if } F'(V) = -(\gamma - r) \\ = F(V) - W^m(V) & \text{if } F'(V) < -(\gamma - r). \end{cases} \tag{96}$$

As in the baseline, it follows that $\lim_{t\to\infty} V_t = V^B(q)$. Recall $V^B(q) = 0$. 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 (96), this requires that $W^s(V) > 0$ for $V > 0$ and therefore $F'(V) \leq -(\gamma - r)$ for $V > 0$.

Suppose that $F'(V) < -(\gamma - r)$ for $V > 0$, so $W^s(V) = F(V) - W^m(V)$. Inserting this expression into (95) and simplifying leads to the ordinary differential equation

$$(\gamma+\delta)F(V) = \max_{a,W^m} \Big\{\mu - \frac{\phi a^2}{2} - \lambda F(V) + \delta F(0) + F'(V)\big((\gamma+\lambda+\delta)V - F(V) + W^m\big)\Big\}, \tag{97}$$

whereby $a = W^m/\phi$.

Using the envelope theorem to totally differentiate the HJB equation (97) (under the optimal controls) with respect to V yields

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

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 (93)). 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 $\hat{V} > 0$ with $F'(\hat{V}) = -(\gamma - r)$, then there exists $0 < V' < \hat{V}$ with $F'(V') > -(\gamma - r)$, a contradiction. As a result, $F'(V) < -(\gamma - r)$ for all $V > 0$.

The maximization in (97) with respect to monitoring effort yields

$$a(V) = \frac{F(V) - F'(V)V + F'(V)\phi}{\phi}, \tag{98}$$

which is (40). 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 (96) that $F'(0) \geq -(\gamma - r)$. As $F'(V) < -(\gamma - r)$ for all $V > 0$, it follows that $\lim_{V\to 0} F'(V) = -(\gamma - r)$. An alternative way to derive this boundary condition is as follows. Comparing (19) with (97), 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 (97) as V approaches zero.

Because

$$dW_t^s = (\gamma + \lambda_t)W_t^s dt - dC_t^s,$$

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 $V_0 = \kappa q$.