

Screening and Monitoring of Corporate Loans*

Sebastian Gryglewicz[†] Simon Mayer[‡] Erwan Morellec[§]

October 17, 2022

Abstract

How much of a loan should a lender dynamically retain and how does retention affect loan performance? We address these questions in a model in which a lender originates loans that it can sell to investors. The lender reduces default risk through screening at origination and monitoring after origination, but is subject to moral hazard. We show that the optimal lender-investor contract can be implemented by having the lender sell its stake in the loan over time, rationalizing loan sales after origination, and use the model to generate predictions linking loan characteristics to initial retention, sales dynamics, and loan performance.

Keywords: dynamic agency, screening, monitoring, loan sales, loan performance.

JEL Classification: G21, G32.

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

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

[‡]HEC Paris. E-mail: mayer@hec.fr.

[§]EPF Lausanne, Swiss Finance Institute, and CEPR. E-mail: erwan.morellec@epfl.ch.

Over the past 20 years, outstanding corporate debt in the U.S. has nearly tripled. This increase has been fueled by the emergence of an active and liquid secondary market for corporate loans (Saunders, Spina, Steffen, and Streit (2021)), in particular for syndicated corporate loans,¹ as well as by the growth of collateralized loan obligations (CLOs),² in which a broad array of nonbank financial institutions invest (Cordell, Roberts, and Schwert (2021)).³ These developments have given loan originators the possibility to reduce their exposure to borrowers' default risk by selling their stake over the loan's life (Blickle, Fleckenstein, Hillenbrand, and Saunders (2022)). As a result, concerns have been expressed that problems in the corporate debt markets are building up in a similar way as they did in the run-up to the subprime mortgage market crisis.

A key difference between mortgages and corporate loans is that in addition to the screening that takes place prior to origination, lenders (e.g. banks) reduce risk and add value to corporate loans through frequent monitoring over the life of the loans. However, if a lender sells (part of) the loans it has originated, it may not have sufficient incentives to screen and monitor borrowers (Pennacchi (1988) or Gorton and Pennacchi (1995)). While loan sales and their consequences for the mortgage market have been the subject of considerable research, much less is known about the relation between *skin in the game* (i.e., the share retained by originators) and screening, monitoring, and default risk in corporate loan markets.

In this paper, we develop a tractable, unifying framework to study optimal incentive provision for screening and monitoring in credit markets. Our model applies to corporate loans and, in particular, to syndicated lending, but is sufficiently general to apply to other markets. We then derive implications for the dynamically optimal originator share and its

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

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

³As documented for instance in Benmelech, Dlugosz, and Ivashina (2012), the securitization of corporate loans is fundamentally different from the securitization of other asset classes. Corporate loans are significantly larger than mortgages and are typically syndicated. The bank that originated the loan generally retains a fraction of the loan on its balance sheet. Fractions of the same underlying loan are simultaneously held by CLOs as well as by other institutional investors and banks. In addition, each loan included in CLOs is rated.

effects on credit risk. This allows us to (i) shed light on existing empirical findings and (ii) generate new implications regarding the effects of loan and lender characteristics on screening, monitoring, and default risk.

We start our analysis by formulating a dynamic agency model in which a lender (the agent, e.g., the lead bank in a loan syndicate) originates a loan and sells this loan to competitive investors (the principal; e.g., other financial institutions in the syndicate). The loan generates coupon payments up to default or maturity. When originating the loan, the lender may undertake costly screening effort that results in a lower expected default rate. It may also monitor the loan at a cost afterward to further reduce default risk. The loan default intensity is thus endogenous and decreases with the agent's screening and monitoring efforts. Because screening and monitoring are not observable, there is moral hazard and the lender's screening and monitoring incentives pin down the respective effort levels. The lender has a lower valuation for the loan than investors due to a higher discount rate arising from, e.g., regulatory or capital constraints. There are therefore gains from selling the loan to investors. Loan sales however reduce the lender's exposure to loan performance and undermine its incentives, thereby increasing credit risk and reducing the loan value.

We derive the optimal contract between the lender (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 lender and investors. Incentive provision for screening and monitoring requires exposing the lender to loan performance. As the lender is protected by limited liability, this is achieved by delaying its payouts so that it loses its expected future payments upon default. Delaying payments, however, is costly due to the lender's higher discount rate. Based on this trade-off, the paper derives an incentive compatible contract that maximizes total surplus. This contract takes a simple form: The lender retains a share of the loan at origination that it gradually sells over time. In addition, under this optimal contract, the selloff speed decreases over time, so most of the loan sales occur relatively shortly after origination, in line with observed practice ([Blickle et al., 2022](#)).

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 that the agent's incentives by means of delayed payouts are especially strong at origination and decrease over time. Accordingly, monitoring incentives decrease, and hence default risk increases over time. To achieve this reduction in deferred compensation and monitoring incentives, the optimal contract mandates smooth, time-decreasing payments to the agent. Therefore, the optimal contract can be implemented by requiring the loan originator (the lead bank in the case of syndicated loans) to retain a stake in the loan that it gradually sells to investors.

The model predicts that the loan originator initially retains a significant fraction of the loan, in line with the evidence in [Benmelech et al. \(2012\)](#) and [Gustafson, Ivanov, and Meisenzahl \(2021\)](#). Initial retention is lower when intrinsic (pre-screening) credit risk is high (due, e.g. to a risky collateral), when the cost of screening is high (due to, e.g., a higher fraction of soft information), when loan maturity is short, or when the originator's cost of capital is large. In addition, and also in line with the findings in [Blickle et al. \(2022\)](#), our model predicts (*i*) that the originator's share in the loan should decrease over time and the originator may sell (nearly) all of its stake shortly after origination and that (*ii*) the selloff speed is greater when intrinsic credit risk or the lender's cost of capital are larger.

We also show that screening and monitoring are complements, in that an increase in the cost of screening or monitoring leads to a decrease in the optimal levels of both screening and monitoring. The reason is that when, for instance, monitoring is costly, it is optimal to reduce monitoring incentives. As screening and monitoring incentives exhibit synergies, the reduction in monitoring incentives reduces screening incentives. In addition, both screening and monitoring are negatively associated with credit risk, in line with the evidence on

screening in [Ivashina \(2009\)](#) and on monitoring in [Wang and Xia \(2014\)](#) and [Gustafson et al. \(2021\)](#). Interestingly, our results point to a two-way causality: Not only do screening and monitoring reduce credit risk, but intrinsic credit risk (pre-screening) also dampens monitoring and screening efforts. Through this mechanism, our model provides a rationale for the segmentation observed in credit markets, whereby banks (lenders) that exert high (low) screening and monitoring typically finance high (low) quality borrowers.⁴

An important question is whether the share of the originator (the lead arranger’s share in a loan syndicate) can proxy for screening or monitoring incentives and therefore predict loan performance. We show that while initial originator retention is monotonic in the cost of screening and the level of screening effort, it is non-monotonic in the cost of monitoring and the level of monitoring effort. This suggests that the *initial* share of the originator can serve as a proxy for screening effort, but not for monitoring effort because subsequent loan sales imply that monitoring incentives decrease over time. We additionally show that while selloff speed is monotonic in the cost of monitoring and the level of monitoring effort, it is non-monotonic in the cost of screening and the level of screening effort. The non-monotonic relationships between selloff speed and screening as well as between initial retention and monitoring imply that neither initial retention nor a measure of selloff speed can (on their own) proxy for both screening and monitoring, which helps explain the finding of [Blickle et al. \(2022\)](#) that initial retention or selloff speed need not predict loan performance.

Next, we study how debt maturity affects the incentives to screen and monitor. A shorter loan maturity reduces the length of time over which the lender is exposed to loan performance, which weakens its incentives to screen and raises credit risk. To counteract this effect, the optimal contract front-loads and concentrates incentives in the early stages of the contract, which implies higher monitoring incentives initially. Relative to long maturity debt, short maturity debt thus features less screening but more monitoring early on in the lending period, which is implemented by lowering initial retention and increasing the selloff speed. However, because monitoring has less persistent effects than screening and the initially high-

⁴Relatedly, [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.

powered monitoring incentives taper off over time as the lender sells off her stake, we find that loans with shorter maturity have (everything else equal) higher default risk in our model with endogenous default intensity.

One way for loan originators to reduce their skin in the game is to use securitization, for example by including CLOs in the loan syndicate. As discussed in [Daley, Green, and Vanasco \(2020\)](#), the development of markets for securitized products has been facilitated in part by credit rating agencies, “which allow issuers access to a large pool of investors who would otherwise have perceived these securities as opaque and complex.” Indeed, a feature that CLOs share is that each loan included in the deal gets rated. By providing information about initial credit quality, credit ratings at origination generate screening incentives, as lax screening induces a low rating, but do not generate incentives for monitoring which occurs after origination. This implies that screening incentives no longer need to be provided so that loans that are rated are characterized both by lower initial retention by originators and by weaker monitoring incentives. That is, the model predicts that monitoring should be less intensive for syndicated loans with CLOs.

In some applications of credit securitization (e.g., for mortgages), screening and monitoring of loans are generally undertaken by separate entities: An originator responsible for screening and a servicing company in charge of monitoring ([Demiroglu and James \(2012\)](#)). In other settings (e.g., for corporate loans), they are typically undertaken by the same entity. To understand whether bundling affects incentives and credit risk, we consider a model variant in which two otherwise identical agents, called screener and monitor, respectively screen and monitor loans and are both subject to moral hazard. For the screener and monitor to have adequate incentives, they must retain a stake in the securitized loan. However, raising one agent’s incentives and stake in the loan necessarily limits the other agent’s stake and incentives, leading to negative spillovers between the monitor’s and the screener’s incentives. By contrast, when screening and monitoring are undertaken by the same agent, there are positive spillovers between screening and monitoring incentives, making it optimal to bundle the two tasks to exploit these incentive synergies and reduce credit risk.

The model predicts relatively low levels of screening and monitoring in credit markets where these two tasks are separated, as is common for mortgages. According to our model, bundling is particularly beneficial for high quality borrowers—providing a rationale for banks’ focus on this segment of credit markets—and when the benefits of screening and monitoring are high relative to their cost, which is the case for corporate loans whose default risk strongly depends on screening and monitoring.

Our paper relates to the large banking literature on screening and monitoring. Most models in this literature are static; see e.g. [Diamond \(1984\)](#), [Gorton and Pennacchi \(1995\)](#), [Holmstrom \(1989\)](#), or [Parlour and Plantin \(2008\)](#). As a result, they do not explicitly distinguish between monitoring after loan origination and screening of loans at origination and cannot investigate the dynamics of incentives and loan sales and their effects on credit risk. Following early contributions by [Sufi \(2007\)](#) and [Ivashina \(2009\)](#), a growing empirical literature examines the effects of the loan stake of the lead arranger in syndicated loans on screening and monitoring (see e.g. [Benmelech et al. \(2012\)](#), [Wang and Xia \(2014\)](#), [Bord and Santos \(2015\)](#), or [Gustafson et al. \(2021\)](#)). Most of these studies proxy skin in the game by the originator’s initial stake in the loan. This literature has recently focused on loan sales after origination and their effects on incentives and credit risk (see e.g. [Lee, Liu, and Stebunovs \(2022\)](#) or [Blickle et al. \(2022\)](#)).

Our paper contributes to this literature mainly in two ways. First, it highlights the key role of the lender’s stake for screening and monitoring incentives, and rationalizes sales after origination as part of an optimal contract between originators and outside investors. Second, it sheds light on the complex relationship between screening and monitoring and the originator’s stake. In particular, it demonstrates that both initial retention and selloff speed determine incentives. Notably, our results have direct implications for the empirical measurement of screening and monitoring as well as their cost which are typically not observed by empiricists. Our findings suggest that initial retention by the loan originator is a good measure for screening at origination but not for monitoring after origination. Instead, empirical measures for monitoring should take into account the selloff dynamics after origi-

nation. Notably, monitoring should increase with the lead bank’s incentives, as captured by the contemporaneous lead share, in line with the evidence in [Gustafson et al. \(2021\)](#).

From a modeling perspective, our paper builds on the literature that studies dynamic contracts in continuous time, starting with [DeMarzo and Sannikov \(2006\)](#) and [Biais, Mariotti, Plantin, and Rochet \(2007\)](#). In this literature, [Piskorski and Westerfield \(2016\)](#), [Malenko \(2019\)](#), [Orlov \(2022\)](#), and [Gryglewicz and Mayer \(2022\)](#) analyze incentive provision with optimal dynamic contracts and monitoring. [Halac and Prat \(2016\)](#), [Varas, Marinovic, and Skrzypacz \(2020\)](#), and [Hu and Varas \(2021\)](#) characterize optimal monitoring in dynamic settings but do not focus on optimal contracts. In a related paper, [Hartman-Glaser, Piskorski, and Tchisty \(2012\)](#) study optimal securitization and screening of mortgages under moral hazard. In their model, the optimal contract features a single payout to the agent when sufficient time has elapsed after origination. [Malamud, Rui, and Whinston \(2013\)](#) and [Hoffmann, Inderst, and Opp \(2021\)](#) generalize [Hartman-Glaser et al. \(2012\)](#) by allowing for more general preferences and sources of uncertainty, respectively. [Hoffmann, Inderst, and Opp \(2022\)](#) study optimal regulation of compensation in a similar framework.

Our paper advances this literature mainly in two ways. First, unlike ours, these papers do not model screening and monitoring and, as a result, cannot study optimal dynamic incentive provision in corporate loans. Second, we show that the combination of screening and monitoring moral hazard implies that the optimal contract can be implemented by requiring the lender to retain a time decreasing stake in the loan, a result that does not obtain in [Hartman-Glaser et al. \(2012\)](#) or [Hoffmann et al. \(2021, 2022\)](#). 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 loans it retains. This implementation of the optimal contract rationalizes recent empirical findings (such as those in [Gustafson et al. \(2021\)](#) or [Blickle et al. \(2022\)](#)) and allows us to generate unique and novel predictions on the effects of loan characteristics and moral hazard on the lender’s initial retention level as well as the sell-off dynamics. Existing theories cannot generate such predictions.

1 Model setup

Time t is continuous and defined over $[0, \infty)$. A lender (the agent or “she”) originates a loan that can be sold to competitive outside investors (the principal or “they”). In the baseline model, we assume for simplicity that the loan has infinite maturity. Section 4 extends the baseline model by introducing loans with finite maturity and shows that the model’s key implications are robust to the level of maturity. The loan promises a constant flow payoff (coupon payments) normalized to 1 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 (endogenous) intensity $\lambda_t > 0$ at time t , where $\tau := \inf\{t \geq 0 : dN_t = 1\}$. That is, over a short period of time $[t, t + dt)$, the loan defaults with probability $\mathbb{E}dN_t = \lambda_t dt$.

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

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

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

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

A high (low) value of $\bar{\tau} := \bar{\tau}_0$ at time $t = 0$ corresponds to low (high) credit risk.

Screening entails a cost $\frac{1}{2}\kappa q^2$ at time zero. Monitoring entails a flow cost $\frac{1}{2}\phi a_t^2$ 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 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 lender decides to shirk on either task, the loan will have a higher default rate. Also notice that while they both reduce default risk, monitoring and screening differ in two ways. First, screening occurs once when the loan is originated at time $t = 0$, whereas monitoring occurs frequently, specifically at any point in time $t \geq 0$ up to default. Second, the effect of screening is more persistent than that of monitoring, where we consider for tractability that monitoring a_t has a purely transitory impact.

Both the principal (e.g., investors in the syndicate) and the agent (e.g., the lead bank) are risk neutral.⁵ The principal discounts cash flows at rate $r \geq 0$. The agent is more impatient and discounts cash flows at rate $\gamma > r$. The difference in discount rates may reflect the credit constraints or regulatory capital requirements, as in DeMarzo and Duffie (1999), or differences in financial constraints or risk-aversion, as in DeMarzo and Sannikov (2006).

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

Both the principal and the agent are protected by limited liability. That is, the continuation payoff of the principal and the agent from following the contract \mathcal{C} must at any time exceed their outside option, which we normalize to zero. Finally, while we do not impose

⁵Alternatively, one can interpret payoffs and probabilities as evaluated under the risk-neutral measure, in which case the default probability λ_t can be seen is the risk-neutral or “risk-adjusted” default probability.

any explicit constraints on the transfers dC_t , we show later that optimal transfers satisfy $dC_t \geq 0$ for $t > 0$, so the agent, i.e., lender, receives positive payouts $dC_t \geq 0$ over each time interval $[t, t + dt]$ after time zero.

Contracting problem

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

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

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

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

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

$$P_t = \mathbb{E}_t \left[\int_t^\tau e^{-r(s-t)} (1 - c_s) ds \right] = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} (1 - c_s) ds. \quad (4)$$

In equation (4), the second equality integrates the default intensity λ_s over the relevant time

interval. The lender receives P_0 dollars at time $t = 0^-$ from selling the security to investors, in that $dC_{0^-} = P_0$. As outside investors are competitive, the lender can extract all the surplus and therefore chooses the security that maximizes total initial surplus $F_{0^-} := W_{0^-} + P_0$ at time $t = 0^-$. That is, the lender solves

$$\max_{\mathcal{C}} F_{0^-}, \quad (5)$$

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

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

where the second equality integrates the default intensity λ_s over the relevant time interval. W_t is the present value of the future payments to the lender, adjusted for the cost of effort. As such, W_t captures the value of the lender's 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 lender's incentives for screening and monitoring and, hence, the resulting 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 lender 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 to 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 $\lambda_t dt = (\Lambda - a_t - q)dt$. Thus, the agent's optimal monitoring effort is

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

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

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

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

While monitoring a_t impacts the default intensity λ_t at a single point in time t , screening q affects all future default intensities $\{\lambda_t\}_{t \geq 0}$ and thus 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 q to maximize W_{0-} which is the value of her claim after screening is chosen, $W_0(q)$, net of the screening effort cost, $\frac{\kappa q^2}{2}$:

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

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

$$V_t = \frac{\partial}{\partial q} W_t(q). \quad (9)$$

We can use V_0 to write the first-order condition solving (8) for the optimal screening effort:

$$q = \frac{V_0}{\kappa}. \quad (10)$$

V_t captures the agent's screening incentives at time t and, because screening effort is chosen at time $t = 0^-$, V_0 determines the amount of screening q exerted by the agent. Lemma 1 below derives a condition such that the first-order approach is valid. Under that condition, equation (10) describes incentive compatibility for screening effort, in that $q = \hat{q} = \frac{V_0}{\kappa}$.

While the initial value V_0 determines screening effort, the optimal contract will depend on the whole path of V_t beyond $t = 0$. To characterize V_t , we differentiate the integral representation of W_t in equation (6) under the optimal control a_t . When differentiating W_t , we can ignore the effect on a_t due to the envelope theorem. Note also that because screening effort q is neither observable nor contractible, an unobserved change in screening effort q cannot affect contracted flow payments c_t . Accounting only for the direct effect of q on W_t , we get that⁶

$$V_t = \int_t^\infty (s-t)e^{-\gamma(s-t)-\int_t^s \lambda_u du} \left(c_s - \frac{\phi a_s^2}{2} \right) ds = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} W_s ds. \quad (11)$$

Equation (11) reveals a simple interpretation of V_t and of screening incentives in our model. Specifically, as a derivative of the lender's continuation value with respect q , which is a persistent component of the discount rate, V_t is closely related to the notion of *duration*. To obtain the duration of the lender's exposure to the loan, one needs to scale V_t by the value of the exposure, that is, the duration measured in units of time is equal to $D_t = \frac{V_t}{W_t}$. It follows that screening incentives V_t are equal to the product of the duration and value of the lender's exposure, i.e., $V_t = D_t W_t$. The duration D_t measures how long it takes on average for the lender to receive payments from the loan (see the middle part of (11) in which dates are

⁶To see that the last part of the equation holds, note that

$$\begin{aligned} \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} W_s ds &= \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} \int_s^\infty e^{-\gamma(v-s)-\int_s^v \lambda_u du} \left(c_v - \frac{\phi a_v^2}{2} \right) dv ds \\ &= \int_t^\infty \int_s^\infty e^{-\gamma(v-t)-\int_t^v \lambda_u du} \left(c_v - \frac{\phi a_v^2}{2} \right) dv ds = \int_t^\infty \int_t^v e^{-\gamma(v-t)-\int_t^v \lambda_u du} \left(c_v - \frac{\phi a_v^2}{2} \right) ds dv \\ &= \int_t^\infty (v-t)e^{-\gamma(v-t)-\int_t^v \lambda_u du} \left(c_v - \frac{\phi a_v^2}{2} \right) dv, \end{aligned}$$

where the first line uses (6) and the second line changes the order of integration. An alternative derivation of (11) is provided in the proof of Proposition 2 in Appendix C.

weighed by payments). If the duration is high, payments accrue over a long period of time, and the impact of permanent changes in default risk via q is large. At the same time, the timing of payments to the lender affects W_t . Due to discounting and relative impatience of the lender, late payments generate less value and provide less screening incentives than early payments. Thus the decomposition of V_t as a product of W_t and D_t captures the intuition that screening incentives are the strongest if the exposure to the loan is large and with high duration. In general, late payments increase duration but decrease value. The maximization of screening incentives must therefore resolve the tension between duration and value.

Equation (11) also shows that monitoring incentives by means of deferred payouts W_s (for $s \geq t$) pin down screening incentives V_t . That is, screening and monitoring incentives are closely linked and interact with each other. Higher W_t exposes the agent's compensation more strongly to loan performance and therefore motivates screening. In addition, higher W_t boosts monitoring a_t , which delays default and strengthens screening incentives.

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

$$\dot{W}_t := \frac{dW_t}{dt} = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - c_t. \quad (12)$$

Similarly, differentiating V_t in (11) with respect to time t , we obtain the dynamics of V_t :

$$\dot{V}_t := \frac{dV_t}{dt} = (\gamma + \lambda_t)V_t - W_t. \quad (13)$$

We close this section by stating some regularity conditions that we impose on the problem.

Lemma 1. *Suppose that the model parameters satisfy*

$$\kappa > \frac{2}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})^2} + \frac{1}{\phi(r + \Lambda - \bar{a} - \bar{q})^2(\gamma + \Lambda - \bar{a} - \bar{q})^3}. \quad (14)$$

Incentive conditions (7) and (10) hold and uniquely pin down monitoring and screening efforts. Incentive conditions (7) and (10) are sufficient and the first-order approach is valid.

Throughout the paper, we assume that condition (14) in Lemma 1 is met. In addition, we assume that

$$\kappa > \frac{\phi \bar{a}}{\bar{q}(\gamma + \Lambda - \bar{a} - \bar{q})}, \quad (15)$$

which is needed in the proof of Proposition 2.

2.2 Optimal contract

2.2.1 Benchmark: observable and contractible screening

To highlight the differences between monitoring and screening incentives more thoroughly, we start by studying the “second-best” benchmark in which screening is not subject to moral hazard, in that q is publicly observable and contractible. To solve the model under this benchmark, we first fix the screening level q . We conjecture (and verify) that the optimal contract is stationary and features constant flow payments to the manager $c_t = c = c^B(q) > 0$ 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 (12) yields

$$c = (\gamma + \Lambda - a - q)W + \frac{\phi a^2}{2}. \quad (16)$$

Equation (16) implies a one-to-one mapping between c and W . As a result, controlling c is equivalent to controlling W and we can treat W as a choice variable instead of c . Given screening effort q and constant monitoring effort a , the default rate is constant and equal to $\Lambda - a - q$, and the price of the security becomes:

$$P^B(q) = \frac{1 - c}{r + \Lambda - a - q}. \quad (17)$$

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

Next, note that given a screening level q , the optimal monitoring effort a (and equivalently optimal deferred compensation $W = \phi a$) is chosen to maximize total surplus after screening

is chosen, $F^B(q) = P^B(q) + W$. Using equations (16) and (17), we get that the lender solves

$$F^B(q) = \max_{W \in [0, F^B(q)]} \left(\underbrace{\frac{1}{r + \Lambda - a - q}}_{\text{Market value}} - \underbrace{\frac{(\gamma - r)W}{r + \Lambda - a - q}}_{\text{Agency cost}} - \underbrace{\frac{\frac{\phi a^2}{2}}{r + \Lambda - a - q}}_{\text{Monitoring cost}} \right), \quad (18)$$

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 (18) shows that the surplus $F^B(q)$ consists of the value of the loan repayments minus agency and direct cost of monitoring. Because the lender is subject to moral hazard, it must retain a stake W , which generates agency costs due to its relative impatience, $\gamma > r$. The maximization problem in (18) 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 (11), we can also calculate

$$V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}. \quad (19)$$

Equation (19) characterizes the agent's screening incentives under the second-best solution and plays an important role in the solution with non-contractible screening.

Finally, we can optimize $F^B(q)$ over q to determine the optimal screening level in this second-best benchmark: $q^B = \arg \max_{q \in [0, \bar{q}]} \left(F^B(q) - \frac{\kappa q^2}{2} \right)$, determining second-best monitoring effort $a^B(q^B)$ and deferred payouts $W^B(q^B)$. We summarize our findings in the following proposition.

Proposition 1 (Moral hazard over monitoring). *Suppose that screening effort q is contractible, so that there is no moral hazard with respect to screening. At the optimum, the following holds. 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), (16), and (18). The continuation payoff satisfies $W^B(q) < F^B(q)$. Optimal monitoring effort $a^B(q)$ increases with q . The optimal choice of screening effort, denoted by q^B , maximizes $F^B(q) - \frac{\kappa q^2}{2}$.*

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 as:⁷

$$\begin{aligned} F_t &= \underbrace{\int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} (1 - c_s) ds}_{=P_t} + \underbrace{\int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} \left(c_s - \frac{\phi a_s^2}{2} \right) ds}_{=W_t} \\ &= \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s \right) ds. \end{aligned} \quad (20)$$

As V_t and W_t characterize the agent's incentives and there is no other 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 (20) implies that the total surplus $F(V, W)$ solves:

$$\begin{aligned} rF(V, W) &= \max_{a, c} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V, W) \right. \\ &\quad \left. + F_V(V, W)((\gamma + \lambda)V - W) + F_W(V, W) \left((\gamma + \lambda)W + \frac{\phi a^2}{2} - c \right) \right\}, \end{aligned} \quad (21)$$

where $F_V(V, W) = \frac{\partial F(V, W)}{\partial V}$ and $F_W(V, W) = \frac{\partial F(V, W)}{\partial W}$.⁸ Equation (21) is solved subject to the

⁷For a derivation, take $F_t = P_t + W_t$ in the first line of (20) and take the derivative with respect to time, t , to get

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

The above expression can be integrated over time, t , to arrive at the second line of (20).

⁸For 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

incentive condition (7), the limited liability constraints, and the conjecture that payouts to the lender are smooth, in that $dC = cdt$. Note that it is always possible to stipulate that the lender receives an incremental payout of Δ dollars, which leaves V unchanged but changes W by $-\Delta$ dollars.⁹ That is, controlling payouts to the lender is equivalent to controlling W . As a result, we can formulate the dynamic optimization problem of the lender such that W instead of c enters the HJB equation (21) as a control variable. Optimal payouts to the lender 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 (21), we can rewrite (21) as

$$rF(V) = \max_{a \in [0, \bar{a}], W} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)((\gamma + \lambda)V - W) \right\}, \quad (22)$$

where (with a slight abuse of notation) F is a function of V only and W is a control. Equation (22) is solved subject to the incentive condition for monitoring effort (7), i.e., $W = \phi a$, and the principal's and the agent's limited liability conditions, i.e., $W \in [0, F(V)]$.

Moral hazard over screening and the provision of screening incentives distort the optimal choice of monitoring incentives away from the benchmark with contractible (observable) screening. However, because the optimal contract must provide appropriate screening incentives only at inception at time $t = 0^-$ and the provision of these incentives as well as the distortion of monitoring incentives are costly due to $\gamma > r$, these distortions decrease (20) with respect to time to get

$$\dot{F}_t = (r + \lambda_t)F_t - 1 + \frac{\phi a_t^2}{2} - (\gamma - r)W_t,$$

which becomes (21) after inserting $\dot{F}_t = F_V(V_t, W_t)\dot{V}_t + F_W(V_t, W_t)\dot{W}_t$ and $F_t = F(V_t, W_t)$.

⁹If payouts to the lender are not smooth, then it follows similar to (12) 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, $dW = -\Delta$.

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

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

As time t tends to infinity, the state variable V approaches $V^B(q)$ which is defined in (19). Expressed in terms of the state variable V , equation (22) is solved subject to the boundary condition

$$\lim_{V \rightarrow V^B(q)} F(V) = F^B(q). \quad (23)$$

We 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 $\lim_{t \rightarrow \infty} \dot{V}_t = 0$. 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 $V \in (V^B(q), V_0]$. We also show that the value function is strictly concave.

Having characterized the model solution for $t \geq 0$ and given screening effort q , we are now in a position to endogenize screening effort. Optimal screening effort $q = q^*$ maximizes the initial value of surplus net of the screening cost while satisfying the incentive compatibility condition (10):

$$q^* = \arg \max_{q \in [0, \bar{q}]} \left(F(V_0) - \frac{\kappa q^2}{2} \right) \quad \text{s.t.} \quad V_0 = \kappa q. \quad (24)$$

The following proposition summarizes the properties of the optimal contract.

Proposition 2 (Moral hazard over screening and monitoring). *In optimum, the state variables W_t and V_t are characterized in (6) and (11) respectively, and follow the dynamics (12) and (13) respectively. Furthermore, 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 (22) subject to boundary condition (23).*
2. *Optimal monitoring is characterized by the maximization in (22) subject to (7). Optimal screening effort $q = q^*$ is characterized in (24).*

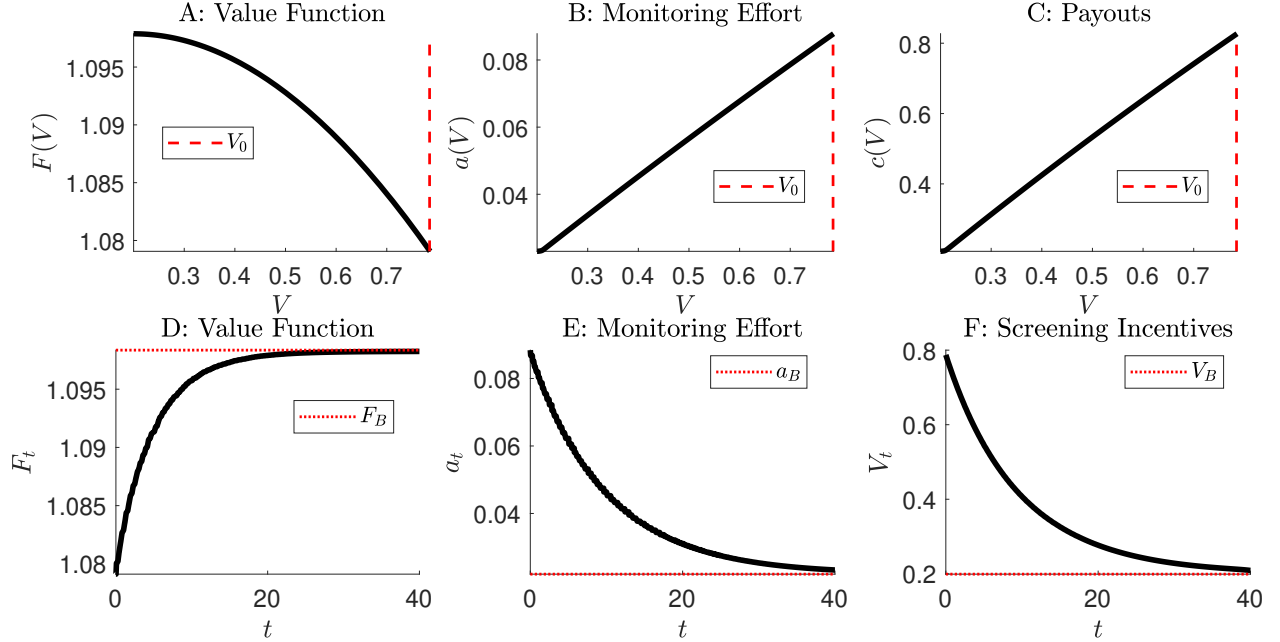


Figure 1: **The quantities characterizing the optimal contract.** In the upper panels, the vertical dashed red line denotes the V_0 . In the lower panels, the horizontal dotted red line denotes the benchmark levels that are attained in the limit $t \rightarrow \infty$.

3. When $q = q^* > 0$, 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., $V_t > V^B(q)$).
4. The value function $F(V)$ strictly decreases in V on $[V^B(q), V_0)$ with $\lim_{V \rightarrow V^B(q)} F'(V) \leq 0$, so that $F'(V) < 0$ for $V > V^B(q)$. The value function is strictly concave.
5. Payouts to the agent are smooth and positive.

Figure 1 provides a numerical example of the optimal contract. For the numerical analysis, we normalize $r = 0$ and $\Lambda = 1$ so that, without monitoring and screening, the expected time to default is $1/\Lambda = 1$ year and the loan has a pre-effort (or intrinsic) value $1/(\Lambda + r) = 1$.¹⁰ 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 $\bar{q} = 0.24$ to satisfy conditions (14) and (15). Our parameter choice implies that the constraints $a_t \leq \bar{a}$ and $q \leq \bar{q}$ never bind. The model's qualitative outcomes are robust to the choice of these parameters.

¹⁰ Λ need not be interpreted as the actual rate of default (absent screening and monitoring), but can rather be seen as risk-adjusted default intensity (i.e., the default intensity under the risk-neutral measure).

The three upper panels of Figure 1 plot total surplus $F(V)$, monitoring $a(V)$, and the agent's flow payouts $c(V)$ as functions of the state variable V . The contract starts at $V = V_0$ and V decreases with time. Observe that flow payouts $c(V)$ to the agent are always positive. Likewise, as $c(V) < 1$ at any time $V \leq V_0$, flow payouts to the principal $1 - c(V)$ are positive too. As V_t is a deterministic function of time (before default), we can represent the evolution of the contract quantities over time. This is done in the lower three panels depicting screening incentives V_t , total surplus F_t , and monitoring effort a_t as functions of time t (for $t < \tau$). As W_t is proportional to a_t by $W_t = \phi a_t$, it is not plotted separately. Observe that V_t , W_t , and a_t decrease over time with a decreasing speed. In contrast, total surplus F_t increases over time. These dynamics of the value function $F_t = F(V_t)$ and monitoring effort $a_t = a(V_t)$ are shaped by the optimal incentive provision for screening. As screening only occurs at time $t = 0$, screening incentives and therefore the agent's exposure to loan performance are front-loaded, thereby inducing a monitoring effort that exceeds the benchmark level $a^B(q^*)$. Intuitively, the provision of screening incentives distorts monitoring incentives upward, which is costly and curbs total surplus. Over time, these distortions taper off, improving total (continuation) surplus F_t which approaches the second-best level in the long run.

3 Incentive provision and implementation

3.1 Dynamics of incentives

We start by analyzing optimal incentives. Optimal monitoring follows from the first-order condition in (22):

$$a(V) = \frac{\overbrace{F(V)}^{\text{Reduction of default risk}} \underbrace{-F'(V)(V + \phi)}_{\substack{\text{Screening} \\ \text{incentives}(>0)}} - \overbrace{(\gamma - r)\phi}^{\text{Agency costs}}}{\underbrace{\phi}_{\text{Physical cost}}} \wedge \frac{F(V)}{\phi}, \quad (25)$$

where $a(V) = \frac{F(V)}{\phi}$ when the limited liability constraint $F(V) = W(V)$ binds and $x \wedge y = \min\{x, y\}$. Optimal monitoring $a(V)$ is determined by several factors. First, monitoring

reduces default risk, but comes at physical costs. Second, monitoring incentives require deferring the agent's payments, which implies that $W > 0$ and is costly due to the discount rate differential, generating agency costs. Third, monitoring incentives are linked to screening incentives V_0 via $V_0 = \int_0^\infty e^{-\gamma t - \int_0^t \lambda_s ds} W_t dt$, in that stronger monitoring incentives at any time $t > 0$ increase screening incentives at time $t = 0$. This effect results from two separate forces: (i) more monitoring a_t reduces the default intensity λ_t , increasing 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 $a^B = a^B(q^*)$; see Figure 1. As screening is only performed at $t = 0$, its benefits for the agent, as captured by V_t in (11), decrease over time within the optimal contract, converging to $V^B = V^B(q^*)$ (see Figure 1). Because the strength of screening and monitoring incentives are linked, monitoring incentives and, hence, monitoring effort also decrease over time, approaching $a^B(q^*)$ in the limit. As a consequence, the instantaneous default rate λ_t 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 $\lim_{a_t \rightarrow \infty} a_t = a^B$.*

To aid in the intuition of the model solution, Figure 2 plots optimal screening and monitoring efforts against the cost parameters ϕ and κ and the baseline default intensity Λ , and the lender's discount rate/cost of capital γ . As monitoring effort a_t changes over time, we plot it at three different times, i.e., $t = 0$, $t = 5$, and $t \rightarrow \infty$, to better capture its dynamics. Panels A, B, E, and F of Figure 2 show that monitoring effort a_t 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.

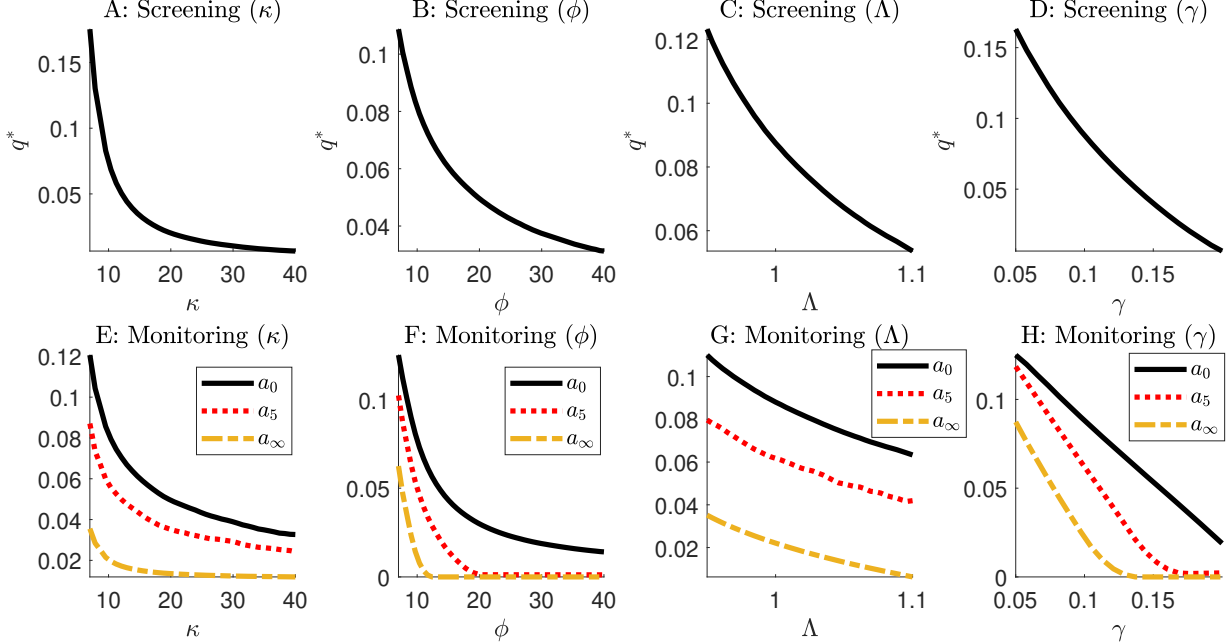


Figure 2: **Comparative Statics.** This figure plots monitoring effort a_t at $t = 0$ (solid black line), at $t = 5$ (dotted red line), and $t \rightarrow \infty$ (dashed yellow line) and screening effort q^* against the parameters ϕ, κ, Λ , and γ . We use our baseline parameters.

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).

Panels C and G of Figure 2 illustrate that a decrease in the quality of the borrower (or in the quality of the loan), as reflected by the higher baseline default intensity 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, lenders who tend to screen or monitor less (e.g., online lenders) finance lower quality borrowers. 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.¹¹

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

3.2 Implementation

This section shows that the optimal contract can be implemented by having the lender retain a time-decreasing share of the loan. At origination, the lender retains a fraction β_0 of the loan and sells a fraction $1 - \beta_0$ to outside investors. After origination at times $t \geq 0$, the lender smoothly sells off its stake β_t so that it decreases over time. That is, the agent owns a fraction β_t of the loan at time t , where β_t is adjusted to provide appropriate incentives W_t .

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

$$L_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} 1 ds, \quad (26)$$

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

Over a short period of time $[t, t + dt]$, the agent receives $\beta_t 1 dt$ in interest payments from the loan. In addition, she sells the loan at rate $-\dot{\beta}_t dt$, which yields trading revenues $-\dot{\beta}_t L_t dt$. Therefore, matching the payoffs of the optimal contract requires that:

$$\beta_t - \dot{\beta}_t L_t = c_t. \quad (27)$$

¹¹A 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 \(2019\)](#). 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.

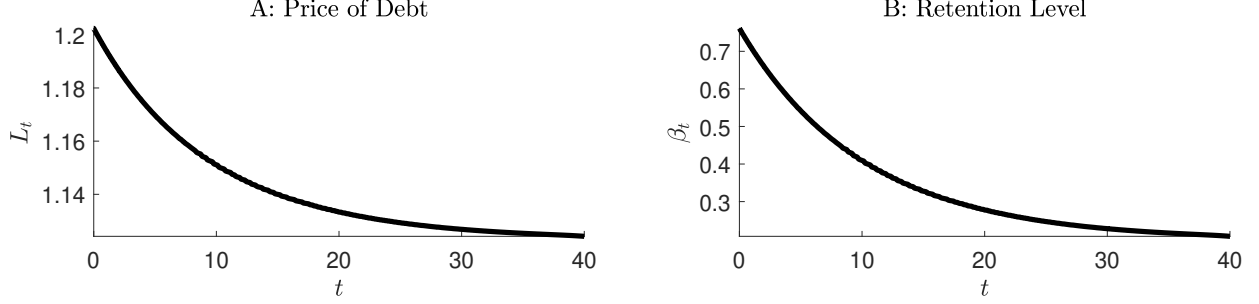


Figure 3: **Implementation of the optimal contract and per unit value of the loan.**

As the HJB equation (22) determines optimal monitoring incentives, and hence optimal deferred compensation $W_t = W(V_t)$, the agent's payouts are implicitly characterized in (12). That is, we can solve (12) to get

$$c_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t > 0. \quad (28)$$

This equation, together with equation (27), implies that

$$\beta_t - \dot{\beta}_t L_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t, \quad (29)$$

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

Figure 3 presents a numerical example of the implementation of the optimal contract and plots the (per-unit) value of the loan and the issuer's stake against time t . As time passes, the agent sells her stake β_t and monitoring incentives decrease, which increases default risk and decreases the (per unit) value of the loan L_t . Also observe that the selloff speed, as, for instance, captured by $-\dot{\beta}_t$, decreases with time t since origination (i.e., β_t is convex and decreasing in t approaching some level β_B). The interpretation is that most of the loan sales occur (relatively) shortly after origination, consistent with (Blickle et al., 2022). 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 loan at time t , whereby a unit stake pays out a flow payoff of 1 dollars until liquidation at time τ and has a competitive time- t price given*

by (26). Over time, the agent sells its stake according to (29).

Finally, it is instructive to discuss the implementation of the optimal contract when there is only one type of moral hazard, i.e., either over screening or monitoring but not both. First, when there is only moral hazard over monitoring (i.e., q is observable and contractible), the solution is characterized in Section 2.2.1, and the optimal contract is time-stationary with constant monitoring $a^B(q) = W^B(q)/\phi$ and constant payouts $c^B(q)$ up to default which can be implemented by having the agent retain a constant share of the loans $\beta^B(q) = c^B(q)$.

Second, Appendix D.3 solves the model when there is no moral hazard over monitoring (i.e., a_t is observable and contractible). As shown in Appendix D.3, the optimal contract can then be implemented by requiring the agent to retain the entire loan until an (endogenous) time τ^0 , at which point the lender sells its entire stake to investors. This implementation maximizes the agent's exposure to loan performance before time τ^0 , while respecting the principal's limited liability. Thus, quite surprisingly, less severe agency conflicts, i.e., removing moral hazard over monitoring, actually increase the lender's optimal initial retention, as optimal initial retention in the baseline model with moral hazard over both tasks is smaller than one. 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 results.

Proposition 4. *When there is no moral hazard over screening, the optimal contract can be implemented by having the agent retain a constant fraction of the loan. When there is no moral hazard over monitoring, the optimal contract can then be implemented by requiring the agent to retain the entire loan until (endogenous) time τ^0 . At time τ^0 , the lender sells its entire stake to the investors.*

3.3 Optimal retention and retention dynamics

The optimal contract between the loan originator and outside investors can be implemented by having the loan originator retain a time-decreasing stake in the loan. As a result, both the initial retention level and the speed at which the lender sells its stake determine the strength of dynamic screening and monitoring incentives. We now study how intrinsic credit risk, the costs of monitoring and screening, and the originator’s cost of capital affect initial retention and selloff dynamics. To this end, the upper three panels of Figure 4 plot the lender’s retention level β_T for $T = 0$ (solid black line), $T = 3$ (dotted red line), and $T \rightarrow \infty$ (dashed yellow line) against κ , ϕ , Λ , and γ . The lower three panels of Figure 4 plot a measure of the selloff speed, $1 - \beta_T/\beta_0$, against κ , ϕ , Λ , and γ . Notice that $1 - \beta_T/\beta_0$ is the fraction of its initial stake that the lender sells up to time T ; thus, if $1 - \beta_T/\beta_0$ is high (low), the lender sells off its initially stake quickly (slowly).

Figure 4 reveals that, as intrinsic credit risk Λ or the lender’s discount rate γ increase, retention decreases and selloff speed increases (see Panels C, D, G, and H), so that the lender’s incentives to screen and monitor decrease, in line with Figure 2. The model, therefore, predicts that originator initially retains a lower fraction of the loan and sells its stake faster when ex-ante credit risk (Λ) is high or when it is more capital-constrained. These results are in line with the findings in [Blickle et al. \(2022\)](#) that lead share sales are positively correlated with the ex-ante riskiness of the loan and the lead arranger’s capital constraints, in [Irani, Iyer, Meisenzahl, and Peydro \(2021\)](#) that less-capitalized banks reduce loan retention, and in [Adelino, Gerardi, and Hartman-Glaser \(2019\)](#) that mortgage quality is positively related to the time to sale for securitized mortgages.¹² Figure 4 also shows that, when ϕ or γ is large, the originator sells nearly its entire share (relatively) shortly after origination as part of the optimal contract, rationalizing the findings of [Blickle et al. \(2022\)](#).

Panels A and E present the effects of the cost of screening κ on retention and selloff speed. Initial retention decreases with κ , however, selloff speed is hump-shaped in κ .¹³ As κ

¹²Although the authors interpret their finding in the context of an adverse selection model (see, e.g., [Daley and Green \(2012\)](#)), our results show that moral hazard generates similar patterns.

¹³These results are robust for a larger range of κ and across different parameter values.

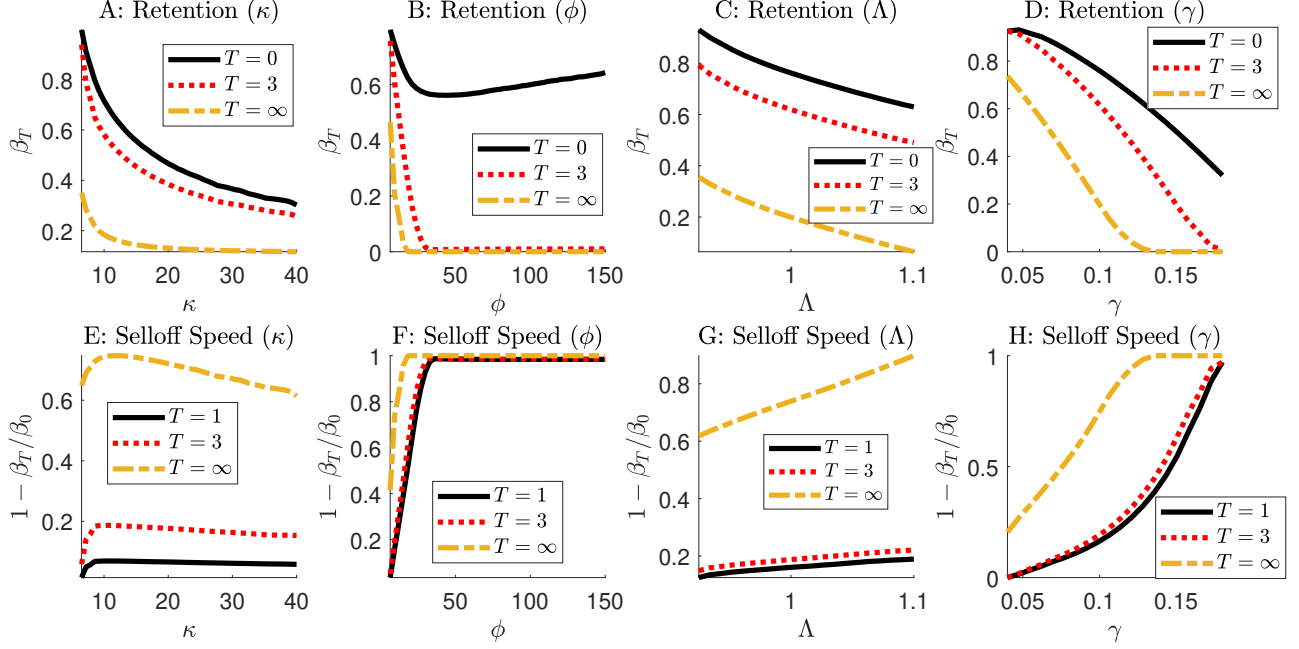


Figure 4: **Retention and dynamics.** The figure plots initial retention and selloff speed as functions of the costs of screening and monitoring κ and ϕ , intrinsic credit quality Λ , and the lender's cost of capital γ .

increases, contracted screening and monitoring efforts decrease, due to the complementarity of screening and monitoring, leading to a decrease in incentives and initial retention. To get some intuition for why selloff speed is the highest for intermediate κ , note that when κ is sufficiently low, moral hazard over screening becomes negligible and the optimal contract only needs to incentivize monitoring. Thus, the contract comes close to that in the benchmark with only monitoring moral hazard and a constant level of retention, that is, a zero selloff speed (see Proposition 4). When κ is sufficiently large and screening is prohibitively costly, there is effectively no moral hazard over screening either as the agent's choice of screening effort tends to zero. Again, in this case, the contract comes close to that in the benchmark with only monitoring moral hazard and a zero selloff speed. Consequently, screening effort, which is monotonically decreasing in κ , can be either increasing or decreasing in selloff speed.

Panels B and F of Figure 4 show the relation between the cost of monitoring ϕ and the levels of retention and selloff speed. Remarkably, in contrast to the effect of κ , initial retention is non-monotonic in ϕ . The intuition for why initial retention is the lowest for

intermediate ϕ is related to the observation that when the cost of monitoring ϕ is sufficiently low or prohibitively high, moral hazard over monitoring becomes negligible and the optimal contract only needs to incentivize screening. According to Proposition 4, absent moral hazard over monitoring, initial retention equals one and selloff occurs only after sufficient time has elapsed. As a consequence, monitoring effort, which is monotonically decreasing in ϕ , can be either increasing or decreasing in initial retention.

The above results have important implications for empirical research on incentives and loan performance. First of all, our model implies that moral hazard in loan screening and monitoring does not generate a simple relation between loan performance and initial retention or selloff speed. As noted above, monitoring effort is non-monotonic in initial retention and screening effort is non-monotonic in selloff speed. Because loan performance depends on both screening and monitoring, these non-monotonic relations help rationalize the finding of [Blickle et al. \(2022\)](#) that initial retention or selloff speed may not predict loan performance.

Instead, the model suggests that screening and monitoring are distinct and that screening and monitoring levels can be separately matched with observables. Notably, we show that while initial retention proxies for screening incentives and effort (in that both initial retention β_0 and screening effort decrease with κ), it does not proxy monitoring incentives and effort (as β_0 is non-monotonic in ϕ but monitoring effort decreases with ϕ). The intuition for this finding is that initial retention is more relevant for screening than for monitoring because screening occurs at origination, while monitoring occurs after origination and thus potentially after the loan originator has sold some of its stake. High initial retention β_0 implies high future retention or high payoffs from loan sales after origination both of which, *ceteris paribus*, raise screening incentives. In contrast, monitoring incentives after time t depend only on the retention level β_t at time t and selloff dynamics after time t , but not directly on β_0 or the loan sales up to time t . Thus, high initial retention, while stimulating screening, may come along with low monitoring incentives when the originator quickly sells off its share after origination.

According to our theory, proxies for monitoring should therefore take into account selloff

dynamics. Indeed, selloff speed increases with ϕ and thus proxies monitoring incentives, whereas it is non-monotonic in κ and so does not proxy screening incentives. Moreover, Panel B of Figure 4 shows that, while initial retention β_0 is U-shaped in ϕ , the lender's retention level β_T at later times T decreases with the cost of monitoring ϕ . Interestingly, and in line with our theory, Gustafson et al. (2021) find that monitoring in a given year is positively related to the lead share in the same year.

3.4 The effects of credit ratings and CLOs

One way to alleviate moral hazard over screening is via a credit rating at origination of the loans. Specifically, consider a setting in which the loan is rated once at origination, i.e., at time $t = 0$ after screening effort has been chosen.¹⁴ For simplicity, we assume that the rating agency perfectly observes the credit quality and reports it truthfully, in that the credit rating is publicly observable and contractible. In our setting, the credit rating reveals the initial credit quality and screening effort q that is chosen at origination.¹⁵ 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. 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 model.

Figure 5 illustrates the effects of credit ratings on outcome variables by plotting the percentage change in monitoring effort (first row), screening effort (second row), and initial retention (third row) at $t = 0$ due to a credit rating. As shown by the figure, the credit

¹⁴This assumption captures the feature of the market that ratings are issued relatively infrequently.

¹⁵Recall that the principal and the agent sign a contract at time $t = 0^-$, just before screening effort is chosen. The credit rating makes q publicly observable and contractible, so one can think of screening and credit rating occurring simultaneously. Another way to think about credit rating is as follows. The rating could also happen after screening effort is chosen: then, investors get their money back (and the contract is renege) if the bank deviates from the promised screening effort, which makes screening effort contractible.

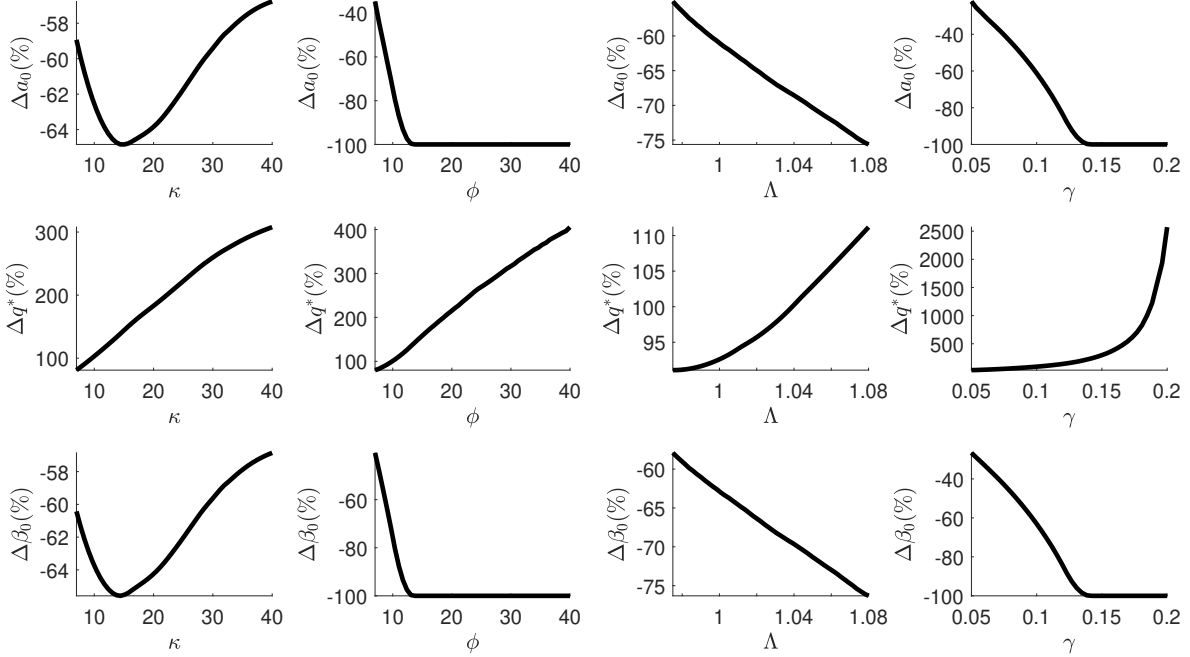


Figure 5: **The effects of credit ratings.** Δy denotes the percentage change in the initial value of the outcome variable y caused by a credit rating, where $y \in \{a_0, q^*, \beta_0\}$. Outcome variables are plotted as functions of the cost of monitoring κ , the cost of screening ϕ , the raw default intensity Λ , and the lender's discount rate γ .

rating increases screening at origination but reduces monitoring a_0 . The reason is that the credit rating increases the agent's incentives to screen loans at origination without requiring increasing its skin in the game. The agent, therefore, requires lower screening incentives through deferred payouts and hence retains a lower share in the loan, leading to lower monitoring incentives. 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 5 shows that under all parameters considered, a credit rating reduces initial retention β_0 . The intuition is that by removing moral hazard over screening, the credit rating allows the bank to reduce its incentives-based exposure to the loan (and eliminate front-loading). In addition, and as shown in Proposition 1, the credit rating affects the optimal retention level and implies that the bank (loan originator) retains a constant stake in the loan. Thus, the credit rating reduces both initial retention and the selloff speed.

A standard way for originators to reduce their share in the loans they originate is to use

securitization, for example, by including CLOs in the syndicate. A salient feature of CLOs is that each loan included in the deal gets rated. Our findings on the effects of credit ratings imply that loans included in a CLO should feature more screening at origination and less monitoring after origination. Moreover, our model predicts that the share retained by the originator should be lower when the originator sells to CLOs.

4 The effects of loan maturity

In our baseline model, loans have infinite maturity. As screening and monitoring efforts have effects of different duration, loan maturity could have different effects on these two tasks. In fact, we do show in this section that loan maturity can have opposing effects on screening and monitoring. To model finite maturity, we follow [Chen, Xu, and Yang \(2021\)](#) and consider that the loan randomly matures with Poisson intensity $\delta > 0$. That is, ignoring default, the expected loan maturity is $1/\delta$. Up to its maturity date, the loan makes coupon payments at rate 1. When the loan matures at t , the firm pays back the face value F_t^δ . That is, at maturity, the game ends, the lender and outside investors exit, and F_t^δ represents their joint terminal payoff. The baseline setting corresponds to the case $\delta = 0$.

With finite maturity loans, the contracting problem is essentially the same as in the baseline model, except that one needs to take into account the impact of finite maturity on the value function and the state variables. With finite maturity, the total continuation surplus satisfies

$$F_t = \int_t^\infty e^{-(r+\delta)(s-t) - \int_t^s \lambda_u du} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s + \delta F_s^\delta \right) ds. \quad (30)$$

This expression differs from that in the baseline model in [\(20\)](#) as the loan matures at rate δ , leading to the terminal payoff F_s^δ when the loan matures at time s . With finite maturity, the agent's screening incentives at time $t = 0$ read

$$V_0 = \int_0^\infty e^{-(\gamma+\delta)t - \int_0^t \lambda_s ds} W_t dt. \quad (31)$$

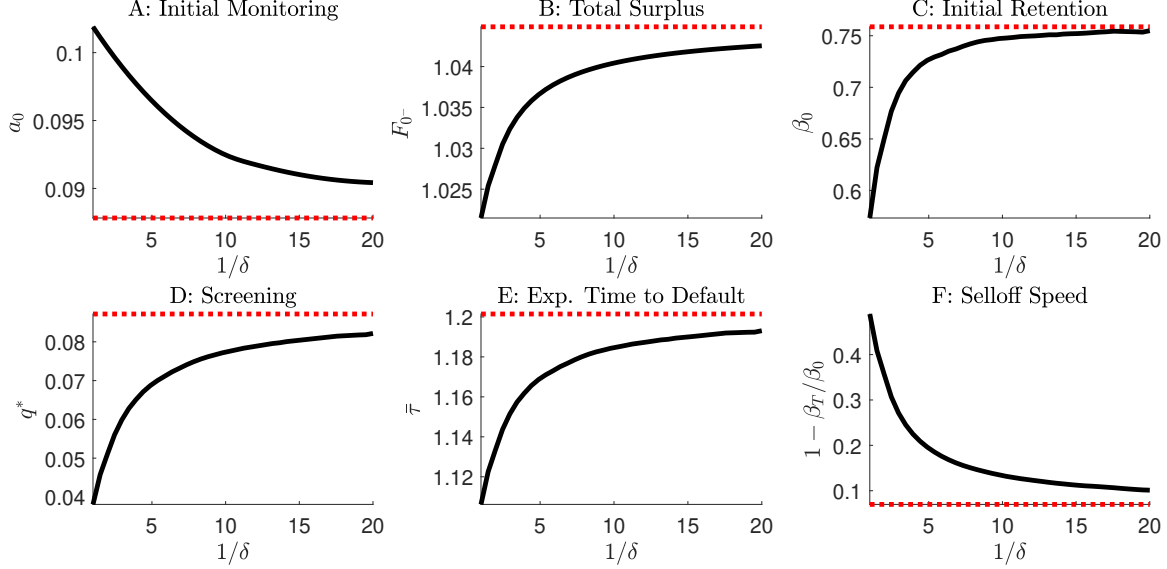


Figure 6: **The effects of debt maturity.** We use our baseline parameters and set $T = 3$ for selloff speed. The dotted red line depicts the outcomes with infinite debt maturity.

At the time of maturity, the lender exits and is no longer exposed to default risk, so its screening incentives fall to zero; thus, the difference between (11) (for $t = 0$) and (31) is that δ augments the discount rate, which reduces screening incentives V_0 . This is also reflected in the law of motion of V_t which becomes

$$\dot{V}_t = (\gamma + \delta + \lambda_t)V_t - W_t. \quad (32)$$

That is, shorter maturity reduces the duration of the lender's claim and thus the lender's exposure to loan performance, thereby undermining screening incentives. In contrast, loan maturity has no direct effect on monitoring incentives, as the impact of monitoring at time t is instantaneous. According to (31), screening incentives V_0 decrease with δ (i.e., increase with loan maturity $1/\delta$). In other words, keeping monitoring effort a_t and the lender's value W_t constant, shorter maturity reduces the duration of the lender's incentives and thus weakens screening incentives. The detailed model description and the remainder of the solution to the model with finite maturity including HJB equation are contained in Appendix D.5.

We numerically solve the model for different loan maturities under our baseline parame-

ters. We find that the contract dynamics in the model with finite maturity are qualitatively similar to those in the baseline model. That is, V decreases over time while monitoring incentives increase and the contract can be implemented by requiring the originator to hold a time-decreasing share of the loan. Moreover, in Appendix D.5.2, we replicate Figures 2 and 4 for finite loan maturity and show that the results remain qualitatively similar. As such, our key findings on retention, selloff dynamics, and screening and monitoring incentives are robust to changing the loan maturity.

Figure 6 plots initial monitoring effort a_0 (which is proportional to the initial value of the lender’s exposure W_0) (Panel A) and screening effort q^* (Panel D) for varying loan maturities. Recall that screening incentives V_0 are a product of the value and the duration of the lender’s exposure. As discussed above, short maturity undermines the lender’s screening incentives by shortening the duration of the lender’s claim and its exposure to loan performance. To counteract this adverse effect, the optimal contract stipulates a higher value of the lender’s initial exposure W_0 for short maturity loans (Panel A). The duration effect dominates, and so screening effort decreases for short maturity loans (Panel D). At the same time, high W_0 generates high initial monitoring incentives and high monitoring effort a_0 for short maturity loans. Therefore, our model predicts relatively low (high) screening but high (low) initial monitoring for corporate loans with a short (long) maturity.

The effects of debt maturity on screening and monitoring feed back into default risk. Notably, Panel E of Figure 6 shows that because monitoring has less persistent effects than screening and the initially high-powered monitoring incentives taper off over time as the lender sells off her stake, loans with shorter maturity have higher default risk (i.e. a lower expected time to default $\bar{\tau}$). Thus, in our model with endogenous default intensity, credit risk decreases as maturity increases (i.e., $\bar{\tau}$ increases with $1/\delta$).¹⁶ Panel B of Figure 6 shows that total surplus increases with debt maturity due to lower agency costs. Our model, therefore,

¹⁶To compare credit risk across different loan maturities on a fair basis, we calculate the expected time to default (at time $t = 0$) conditional on the loans not maturing. That is we use the (inverse) measure of credit risk

$$\bar{\tau} := \int_0^\infty e^{-\int_0^t \lambda_u du} dt$$

which eliminates the effect of maturity on the duration over which the loan is exposed to credit risk.

provides a rationale for the use of long-term debt in the presence of agency frictions at the loan originator level. Figure 6 also plots the initial share retained by the originator and the selloff speed against loan maturity in Panels C and F.¹⁷ The optimal contract implements high initial monitoring for short maturity loans by frontloading the agent’s compensation. Interestingly, this is achieved by increasing the selloff speed for shorter maturity loans (see Panel F), while initial retention β_0 decreases (Panel C).

5 Is it optimal to bundle monitoring and screening?

We have so far assumed that the loan originator is responsible for both screening and monitoring. In practice, screening and monitoring may be undertaken by separate entities. Some securitized loans are serviced by a third-party serving company and, depending on the specific arrangements, servicing can subsume monitoring activities. In these cases, the originator is in charge of screening and the servicer in charge of monitoring. An important question is therefore whether bundling 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; monitoring effort (screening effort) is only and privately observed by the monitor (screener). Appendix D.6 provides the detailed description and solution to the model with separated screening and monitoring tasks. Below, we describe the intuition for the optimal contract, its dynamics, and present numerical results related to key outcome variables under the baseline and this model variant.

Screening and monitoring incentives are provided by having the screener and monitor retain a share of the loan. The screener’s and monitor’s shares add up to one until sufficient time has elapsed and the screener sells off its entire stake at once to investors; the monitor continues to maintain (time-varying) exposure to the loans. Notably, monitoring incentives

¹⁷When calculating the retention level β_t , one must also calculate the market value of debt L_t , as $\dot{\beta}_t L_t + \beta_t \mu = c_t$ holds. We impose “value-matching” when calculating the market value of debt in that the value of debt L_t is the same “just before” and at maturity. This implies $L_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} ds$.

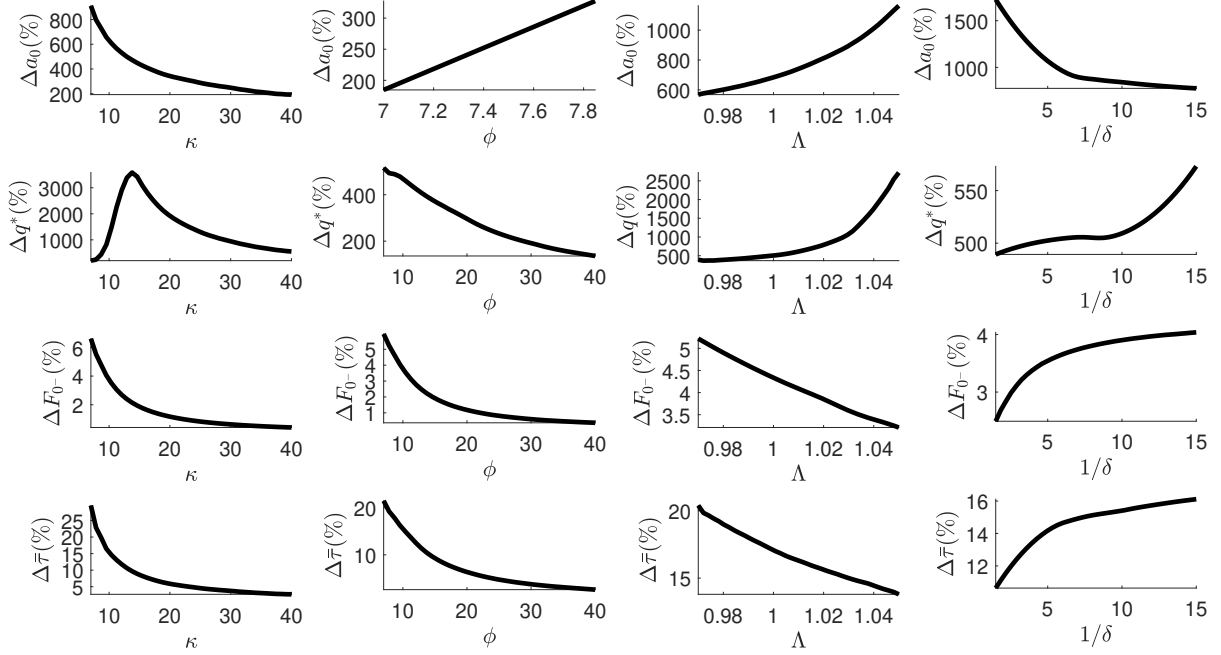


Figure 7: **The effects of bundling screening and monitoring.** Δa_0 denotes the percentage change in monitoring effort at $t = 0$ due to bundling. Δq^* denotes the percentage change in screening effort due to bundling. ΔF_{0^-} denotes the percentage change in total surplus at $t = 0^-$ caused by bundling. $\Delta \bar{T}$ denotes the percentage change in the expected time to default due to bundling. Outcome variables are plotted as functions of the cost of monitoring κ , the cost of screening ϕ , the raw default intensity Λ , and loan maturity $1/\delta$ under the baseline parameters.

(provided to the monitor) have two opposing effects on screening incentives. 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 stake, which, in turn, requires lowering the screener's stake as their shares add up to one. 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 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, increases 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 param-

eters considered, bundling increases (initial) monitoring (i.e., $\Delta a_0 > 0$), screening ($\Delta q^* > 0$), and total surplus ($\Delta F_{0-} > 0$). Our model therefore predicts relatively low levels of monitoring and screening in the mortgage market, where screening and monitoring tasks are often separated (Demiroglu and James, 2012).

Also notice that according to Figure 7, bundling screening and monitoring increases total surplus and reduces credit relatively less, the larger the cost of screening or monitoring, the larger intrinsic credit Λ , or the longer the loan maturity. One interpretation of this result is that when, for instance, monitoring borrowers is difficult after origination in that ϕ is large, bundling of separating and monitoring is less likely to occur. According to our model, bundling is more likely to occur in credit markets in which screening and monitoring are important for credit risk (i.e., the effects of screening/monitoring are large relative to the cost), such as the market for corporate loans.

6 Conclusion

We study a dynamic moral hazard problem in which a lender (e.g., the lead bank in a syndicate) originates a loan to sell it to investors (e.g., other financial institutions in the syndicate). The lender controls the loan’s default risk through screening at origination and monitoring after origination, both of which are subject to moral hazard. Screening and monitoring incentives are provided by exposing the lender to loan performance. As screening occurs only once at the origination of the loan, incentives are front-loaded and stronger shortly after origination. The optimal contract can be implemented by requiring the loan originator to retain a time-decreasing stake in the loan so that its 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. The optimal contract also implies that screening and monitoring decrease with intrinsic (pre-screening) credit risk, suggesting that lenders specializing in financing high-quality borrowers (such as banks) exert higher levels of screening and monitoring.

The unique and novel feature of our paper is that it allows us to analyze how loan and

originator characteristics affect initial retention and subsequent loan sales, thereby rationalizing a number of empirical findings and providing new testable empirical hypotheses. For instance, we show that initial retention decreases while the selloff speed increases with borrowers' intrinsic credit risk, the lender's cost of capital, or loan maturity. Moreover, our model implies that while initial retention increases with the cost of screening, which maps one-to-one to hidden screening effort, it is non-monotonic in the cost of monitoring, which maps one-to-one to hidden monitoring effort. In contrast, the speed at which the lender sells off its stake in the loan increases with the cost of screening, but is non-monotonic in the cost of monitoring. Our model, therefore, suggests that the originator's initial retention can serve as a proxy for screening but not for monitoring incentives, whereas the selloff speed can serve as a proxy for monitoring but not screening incentives.

Our model is simple and general enough that it can be used to analyze a wide range of credit markets. For example, we extend our model to analyze the provision of incentives when screening and monitoring are performed by separate entities, which is often the case for mortgages: An originator that selects loans initially and a servicer that monitors them later. We show that such a separation of monitoring and screening tasks reduces both monitoring and screening effort, thereby increasing credit risk.

Finally, the moral hazard problem we study also has applications in contexts other than credit securitization and syndicated lending. In particular, screening before funding an investment and monitoring afterward is also common in venture capital financing (see [Bernstein, Giroud, and Townsend \(2016\)](#) for evidence on monitoring and [Abuzov \(2022\)](#) for evidence on screening). Our theory could be easily modified to study venture capital financing with moral hazard over screening and monitoring. We leave this for future research.

References

- Abuzov, R. (2022). The impact of venture capital screening. *Working paper, The University of Virginia* (19-14).
- Adelino, M., K. Gerardi, and B. Hartman-Glaser (2019). Are lemons sold first? Dynamic signaling in the mortgage market. *Journal of Financial Economics* 132(1), 1–25.
- Benmelech, E., J. Dlugosz, and V. Ivashina (2012). Securitization without adverse selection: The case of CLOs. *Journal of Financial Economics* 139(2), 452–477.
- Bernstein, S., X. Giroud, and R. R. Townsend (2016). The impact of venture capital monitoring. *The Journal of Finance* 71(4), 1591–1622.
- 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.
- Blickle, K., Q. Fleckenstein, S. Hillenbrand, and A. Saunders (2022). The myth of the lead arranger’s share. *Working paper NYU*.
- Bord, V. M. and J. A. Santos (2015). Does securitization of corporate loans lead to riskier lending? *Journal of Money, Credit and Banking* 47(2-3), 415–444.
- 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. and B. Green (2012). Waiting for news in the market for lemons. *Econometrica* 80(4), 1433–1504.
- 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. 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.
- Gorton, G. B. and G. G. Pennacchi (1995). Banks and loan sales marketing nonmarketable assets. *Journal of Monetary Economics* 35(3), 389–411.
- Gryglewicz, S. and S. Mayer (2022). Dynamic contracting with intermediation: Operational, governance, and financial engineering. *Journal of Finance* (forthcoming).
- Gustafson, M., I. Ivanov, and R. Meisenzahl (2021). Bank monitoring: Evidence from syndicated loans. *Journal of Financial Economics* 139(1), 91–113.
- Halac, M. and A. Prat (2016). Managerial attention and worker performance. *American Economic Review* 106(10), 3104–32.
- Hartman-Glaser, B., T. Piskorski, and A. Tchisty (2012). Optimal securitization with moral hazard. *Journal of Financial Economics* 104(1), 186–202.
- 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.
- Hoffmann, F., R. Inderst, and M. M. Opp (2022). The economics of deferral and clawback requirements. *Journal of Finance* 77(4), 2423–2470.
- 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.
- Irani, R. M., R. Iyer, R. R. Meisenzahl, and J.-L. Peydro (2021). The rise of shadow banking: Evidence from capital regulation. *The Review of Financial Studies* 34(5), 2181–2235.
- Ivashina, V. (2009). Asymmetric information effects on loan spreads. *Journal of Financial Economics* 92(2), 300–319.

- 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.
- Lee, S. J., L. Q. Liu, and V. Stebunovs (2022). Risk-taking spillovers of u.s. monetary policy in the global market for U.S. dollar corporate loans. *Journal of Banking and Finance* 138(2), 105550.
- 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.
- Orlov, D. (2022). Frequent monitoring in dynamic contracts. *Journal of Economic Theory*, 105550.
- Parlour, C. and G. Plantin (2008). Loan sales and relationship banking. *Journal of Finance* 63(3), 1291–1314.
- 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.
- Saunders, A., A. Spina, S. Steffen, and D. Streitz (2021). Corporate loan spreads and economic activity. Technical report.
- Sufi, A. (2007). Information asymmetry and financing arrangements: Evidence from syndicated loans. *Journal of Finance* 62(2), 629–668.
- 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.
- Wang, Y. and H. Xia (2014). Do lenders still monitor when they can securitize loans? *Review of Financial Studies* 27(8), 2354–2391.

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

$$\max_{a_t \in [0, \bar{a}]} \left(a_t W_t - \frac{\phi a_t^2}{2} \right), \quad (\text{A.1})$$

which yields

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

Observe that when optimal monitoring effort is interior and $a_t < \bar{a}$, the above condition simplifies to (7), i.e., $a_t = \frac{W_t}{\phi}$, which is the first order condition to (A.1). The second order condition to (A.1), i.e., $\frac{\partial^2}{\partial a_t^2} \left(a_t W_t - \frac{\phi a_t^2}{2} \right) = -\phi < 0$, is satisfied. Thus, contracted effort level within an incentive compatible contract satisfies $\hat{a}_t = W_t/\phi$.

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

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

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

Now observe that

$$q = \min \left\{ \frac{V_0(q)}{\kappa}, \bar{q} \right\} \quad (\text{A.3})$$

is the unique solution to (A.2) if

$$\frac{\partial^2}{\partial q^2} \left(W_0(q) - \frac{\kappa q^2}{2} \right) = \frac{\partial}{\partial q} V_0(q) - \kappa < 0 \quad (\text{A.4})$$

holds for any $q \in [0, \bar{q}]$, in which case the objective in (A.2) is strictly concave over the entire interval $[0, \bar{q}]$ and the first order approach is valid. When optimal screening effort is interior, condition (A.3) simplifies to (10), i.e., $q = V_0/\kappa$, which is the first order condition to (A.2).

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

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

and note that (A.4) can be rewritten as $Y_0(q) < \kappa$. Next, insert $a_t = W_t(q)/\phi$ into (13) to obtain

$$\dot{V}_t = \frac{dV_t(q)}{dt} = \left(\gamma + \Lambda - \frac{W_t(q)}{\phi} - q \right) V_t(q) - W_t(q), \quad (\text{A.5})$$

bearing in mind $\lambda_t = \Lambda - W_t(q)/\phi - q$. We now differentiate (A.5) with respect to q to obtain

$$\dot{Y}_t = \frac{dY_t(q)}{dt} = (\gamma + \lambda_t)Y_t(q) - 2V_t(q) - \frac{(V_t(q))^2}{\phi}.$$

We can integrate the above ODE over time to obtain

$$Y_t(q) = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} \left(2V_s(q) + \frac{(V_s(q))^2}{\phi} \right) ds \quad (\text{A.6})$$

for all $t \geq 0$. In addition, (11) implies

$$V_t(q) = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} W_s(q) ds \quad (\text{A.7})$$

for all $t \geq 0$. Note now that (owing to $a_t \leq \bar{a}$ and $q \leq \bar{q}$)

$$\lambda_t = \Lambda - a_t - q \geq \Lambda - \bar{a} - \bar{q}. \quad (\text{A.8})$$

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

$$\begin{aligned} W_t &\leq F_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s \right) ds \\ &< \int_t^\infty e^{-(r+\Lambda-\bar{a}-\bar{q})(s-t)} 1 ds = \frac{1}{r + \Lambda - \bar{a} - \bar{q}} =: W^{max} \end{aligned} \quad (\text{A.9})$$

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

Using these two relations (A.8) and (A.9) as well as (A.7), we obtain that

$$\begin{aligned} V_t(q) &< \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} W^{max} ds \leq \int_t^\infty e^{-(\gamma+\Lambda-\bar{a}-\bar{q})(s-t)} W^{max} ds \\ &\leq \frac{W^{max}}{\gamma + \Lambda - \bar{a} - \bar{q}} < \frac{1}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})} \end{aligned} \quad (\text{A.10})$$

Using this inequality (A.10) and the integral representation in (A.6), we obtain that

$$\begin{aligned} Y_t(q) &= \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} \left(2V_s(q) + \frac{(V_s(q))^2}{\phi} \right) ds \\ &\leq \int_t^\infty e^{-(\gamma+\Lambda-\bar{a}-\bar{q})(s-t)} \left(2V_s(q) + \frac{(V_s(q))^2}{\phi} \right) ds \\ &< \frac{1}{(\gamma + \Lambda - \bar{a} - \bar{q})} \left(\frac{2}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})} + \frac{1}{\phi(r + \Lambda - \bar{a} - \bar{q})^2(\gamma + \Lambda - \bar{a} - \bar{q})^2} \right). \end{aligned}$$

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

$$Y_0(q) < \kappa,$$

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

$$\kappa > \frac{2}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})^2} + \frac{1}{\phi(r + \Lambda - \bar{a} - \bar{q})^2(\gamma + \Lambda - \bar{a} - \bar{q})^3}. \quad (\text{A.11})$$

That is, when (A.11) holds, the first order approach is valid and (A.3) or, equivalently, (10) (due to $q < \bar{q}$) pins down screening effort. Note that (A.11) is equivalent to condition (14) (Lemma 1). Also notice that (14) but not per-se necessary.

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

$$F_t = \underbrace{\int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} (1ds - dC_s)}_{=P_t} + \underbrace{\int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} \left(dC_s - \frac{\phi a_s^2}{2} ds \right)}_{=W_t},$$

where

$$P_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} (1ds - dC_s)$$

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

$$dP_t = (r + \lambda_t)P_t dt - 1dt + dC_t \quad (\text{B.12})$$

$$dW_t = (\gamma + \lambda_t)W_t dt + \frac{\phi a_t^2}{2} dt - dC_t. \quad (\text{B.13})$$

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

$$dF_t = dP_t + dW_t \quad (\text{B.14})$$

$$\begin{aligned} &= (r + \lambda_t)P_t dt - 1dt + dC_t + (\gamma + \lambda_t)W_t dt - dC_t + \frac{\phi a_t^2}{2} dt \\ &= (r + \lambda_t) \underbrace{(P_t + W_t)}_{=F_t} dt - 1dt + \frac{\phi a_t^2}{2} dt - (\gamma - r)W_t dt. \end{aligned} \quad (\text{B.15})$$

We can integrate (B.14) over time, t , to get

$$F_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s \right) ds, \quad (\text{B.16})$$

which is (20) from the main text.

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

$$F_0 - \frac{\kappa q^2}{2}, \quad (\text{B.17})$$

where F_0 is characterized in (B.16). Note that it is always possible to stipulate payouts dC_t to the agent, which decreases W_t by amount dC_t . As such, controlling payouts to the agent dC_t is equivalent to controlling the agent's continuation payoff W_t . In the following, we take W_t rather than dC_t as control variable for the dynamic optimization, and we drop the control variable dC_t .

By the dynamic programming principle, total surplus F_t must solve at any time $t > 0$ the HJB equation

$$rF_t = \max_{W_t \in [0, F_t], a_t \geq 0} \left(1 - \frac{\phi a_t^2}{2} - (\gamma - r)W_t + \dot{F}_t - \lambda_t F_t \right),$$

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

$$rF^B(q) = \max_{W \in [0, F^B(q)], a \in [0, \bar{a}]} \left(1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F^B(q) \right) \quad (\text{B.18})$$

subject to the monitoring incentive constraint (7), which can be rewritten as (18).

The maximization in the above HJB equation yields that, if interior, optimal monitoring effort reads

$$a^B(q) = \frac{F^B(q) - \phi(\gamma - r)}{\phi}, \quad (\text{B.19})$$

and the optimal lender continuation value is $W^B(q) = \phi a^B(q)$, due to (7). With a slight abuse of notation, if the above expression for $a^B(q)$ is negative, then optimal monitoring effort $a^B(q)$ is zero. If the above expression for $a^B(q)$ exceeds \bar{a} , then optimal monitoring effort $a^B(q)$ is \bar{a} . Note that the first order condition (B.19) implies $\phi a^B(q) = W^B(q) < F^B(q)$, so the principal's limited liability constraint does not bind in optimum. Since, clearly, $F^B(q)$ increases with q , it follows that $a^B(q)$ increases with q , i.e., $\frac{\partial}{\partial q} a^B(q)$

Optimal monitoring effort implies the instantaneous default probability $\lambda = \lambda^B(q) = \Lambda - q - a^B(q)$. The law of motion (B.12) and $dW_t = 0$ imply then that payouts to the agent take the form $dC_t = c^B(q)dt$ with

$$c^B(q) = (\gamma + \lambda^B(q))W^B(q) + \frac{\phi(a^B(q))^2}{2}. \quad (\text{B.20})$$

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

The objective (B.17) can be rewritten as

$$\max_{q \in [0, \bar{q}]} \left(F^B(q) - \frac{\kappa q^2}{2} \right). \quad (\text{B.21})$$

At time $t = 0$, the agent chooses screening effort $q \in [0, \bar{q}]$ to maximize (B.21), leading to optimal screening effort q^B .

C Proof of Proposition 2

C.1 Preliminaries

To begin with, we derive the dynamics of W_t , i.e., (12), the dynamics of V_t (defined in (9)), and the integral expression (11). Now, recall the definition of W_t in (6) and differentiate (6) with respect to time, t , to obtain

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

which is (12). Using (12), we can write the intermediary's optimization with respect to monitoring effort a_t at time t as

$$\gamma W_t = \max_{a_t \in [0, \bar{a}]} \left(- \underbrace{(\Lambda - a_t - q) W_t}_{=\lambda_t} - \frac{\phi a_t^2}{2} + c_t + \dot{W}_t \right), \quad (\text{C.22})$$

which yields optimal $a_t = \min \left\{ \frac{W_t}{\phi}, \bar{a} \right\}$ (as in (7)) and, as we focus on interior levels, $a_t = W_t/\phi$.

Next, note that because screening effort q is neither observable nor contractible, an unobserved change in screening effort q cannot affect contracted flow payments c_t . We now use the envelope theorem to differentiate both sides of (C.22) under optimal a_t with respect to q so that

$$\gamma V_t = W_t - \lambda_t V_t + \dot{V}_t \iff \dot{V}_t = (\gamma + \lambda_t)V_t - W_t,$$

which is (13) as desired. Note that we used $\frac{\partial}{\partial q} \dot{W}_t = \frac{\partial}{\partial q} \frac{d}{dt} W_t = \frac{d}{dt} \frac{\partial}{\partial q} W_t = \frac{dV_t}{dt} = \dot{V}_t$ as well as $\frac{\partial}{\partial q} \frac{\partial W_t}{\partial a_t} = 0$ (envelope theorem) and $\frac{\partial c_t}{\partial q} = 0$.¹⁸ We can integrate $\dot{V}_t = (\gamma + \lambda_t)V_t - W_t$ over time t to obtain the integral expression (11), that is, $V_t = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} W_s ds$.

The remainder of the proof is split in six parts. Part I characterizes total surplus as a function of the agent's screening incentives $V_t = V$ and shows that in optimum, total surplus (i.e., the value

¹⁸In more detail, note that

$$\frac{d}{dq} W_t = \frac{\partial W_t}{\partial q} + \frac{\partial W_t}{\partial a_t} \frac{\partial a_t}{\partial q} + \frac{\partial W_t}{\partial c_t} \frac{\partial c_t}{\partial q} = \frac{\partial}{\partial q} W_t.$$

as $\frac{\partial W_t}{\partial a_t} = 0$ and $\frac{\partial c_t}{\partial q} = 0$. An alternative derivation (not relying explicitly on envelope theorem) simply rewrites (12) by inserting monitoring incentive compatibility, $a_t = W_t/\phi$, to obtain

$$\dot{W}_t = \left(\gamma + \Lambda - \frac{W_t}{\phi} - q \right) W_t + \frac{W_t^2}{2\phi} - c_t.$$

function $F(V)$) solves the HJB equation (22). Part II demonstrates that $\lim_{t \rightarrow \infty} V_t = V^B(q)$. Part III characterizes the agent's initial choice of optimal screening effort $q = q^*$. Part IV verifies that $\kappa q^* = V_0 > V^B(q^*)$, 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 and is concave. 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 $q \in (0, \bar{q})$. As in the main text, we characterize the solution for $t \geq 0$ given screening effort q , and then determine the optimal screening effort $q = q^*$; unless necessary we do not distinguish notation-wise between q and the optimally chosen screening effort q^* .

We make the following regularity assumption. Throughout, we assume that there exists a unique solution $F(V)$ to the HJB equation (22) which is continuously differentiable. Further, we assume that the second derivative $F''(V)$ exists almost everywhere in the state space $(V^B(q), V_0)$ (i.e., the set of points at which $F'(V)$ is not differentiable is not dense).

C.2 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 (10) holds at time $t = 0$ so that $V_0 = \kappa q$.

The optimal contract maximizes total surplus characterized in (B.16):

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

Note that it is always possible to stipulate payouts dC_t to the agent, which decreases W_t by amount dC_t and leaves V_t unchanged. As such, controlling payouts to the agent dC_t is equivalent to controlling the agent's continuation payoff W_t . In the following, we take W_t rather than dC_t as control variable. Thus, the agent's optimization problem only depends on the state variable V_t summarizing the agent's screening incentives. As a consequence, we can express total surplus as function of V_t , in that $F_t = F(V_t)$. In what follows, we omit time-subscripts whenever possible.

Recall that screening incentives V evolve according to (13), i.e., $\dot{V} = (\gamma + \lambda)V - W$. By the dynamic programming principle, total surplus $F(V)$ solves in any state V the HJB equation

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

which is solved subject to the monitoring incentive constraint (7). Recall that both the principal and the agent are subject to limited liability, so that $W \in [0, F(V)]$ and the principal's payoff

Differentiating both sides with respect to q and using $\frac{\partial c_t}{\partial q} = 0$, we obtain

$$\dot{V}_t = (\gamma + \lambda_t)V_t - W_t - \frac{V_t W_t}{\phi} + \frac{V_t W_t}{\phi},$$

which simplifies to (13), as desired.

$F(V) - W$ satisfies $F(V) - W \in [0, F(V)]$ too. The above HJB equation coincides with (22). The maximization in the above HJB equation yields that, if interior, optimal monitoring effort is

$$a(V) = \frac{F(V) - F'(V)(V + \phi) - (\gamma - r)\phi}{\phi} \wedge W(C), \quad (\text{C.23})$$

which is (25).

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 (13), we can solve for

$$V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}. \quad (\text{C.24})$$

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

$$F(V^B(q)) = F^B(q). \quad (\text{C.25})$$

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

$$\lim_{V \rightarrow V^B(q)} F(V) = F^B(q),$$

as well as $\lim_{V \rightarrow V^B(q)} a(V) = a^B(q)$.

C.3 Part II

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

$$\begin{aligned} \mathcal{L} &= \underbrace{\int_0^\infty e^{-rt - \int_0^t \lambda_u du} \left(1 - (\gamma - r)W_t - \frac{\phi a_t^2}{2} \right) dt}_{=F_0} + \ell \left(\kappa q - \underbrace{\int_0^\infty e^{-\gamma t - \int_0^t \lambda_u du} W_t dt}_{=V_0} \right) \\ &= F_0 + \ell(\kappa q - V_0). \end{aligned} \quad (\text{C.26})$$

where ℓ is the Lagrange multiplier with respect to the screening incentive constraint (10) and $W_t = \phi a_t$ is the effort incentive constraint which we directly insert into the objective function.

Next, we rewrite (B.14) as

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

which can be integrated over time to obtain

$$F_t = \int_t^\infty e^{-r(s-t)} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s - \lambda_s F_s \right) ds. \quad (\text{C.27})$$

Likewise, we can rewrite (13) as

$$dV_t = \gamma V_t dt - W_t dt + \lambda_t V_t dt,$$

which can be integrated over time to get

$$V_t = \int_t^\infty e^{-\gamma(s-t)} (W_s - \lambda_s V_s) ds. \quad (\text{C.28})$$

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

$$\mathcal{L} = \int_0^\infty e^{-rt} \left(1 - (\gamma - r)W_t - \frac{\phi a_t^2}{2} - \lambda_t F_t \right) dt + \ell \left(\kappa q - \int_0^\infty e^{-\gamma t} (W_t - \lambda_t V_t) dt \right). \quad (\text{C.29})$$

We can maximize the Lagrangian point-wise (that is, for each time t) 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:

$$e^{-rt}(F_t - (\gamma - r)\phi - \phi a_t) - \ell e^{-\gamma t}(\phi + V_t) = 0 \quad (\text{C.30})$$

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

$$F_t - (\gamma - r)\phi - \phi a_t - \ell e^{-(\gamma-r)t}(\phi + V_t) = 0. \quad (\text{C.31})$$

We can solve (C.31) for

$$a_t = \frac{F_t - (\gamma - r)\phi - \ell e^{-(\gamma-r)t}(\phi + V_t)}{\phi}. \quad (\text{C.32})$$

Taking the limit $t \rightarrow \infty$ in (C.32) leads to

$$\lim_{t \rightarrow \infty} a_t = \lim_{t \rightarrow \infty} \left(\frac{F_t - (\gamma - r)\phi}{\phi} \right), \quad (\text{C.33})$$

as V_t is bounded (see inequality (A.10) in the proof of Lemma 1 and note that by definition, $V_t \geq 0$).

We conjecture (and verify) that, in the limit $t \rightarrow \infty$, the solution becomes stationary and F_t and a_t become constant, in that

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

for (endogenous) constants \hat{F} and \hat{a} .¹⁹ Note that by (C.33),

$$\hat{a} = \frac{\hat{F} - (\gamma - r)\phi}{\phi}. \quad (\text{C.34})$$

¹⁹Equivalently,

$$\lim_{t \rightarrow \infty} \dot{F}_t = 0 \quad \text{and} \quad \lim_{t \rightarrow \infty} \dot{a}_t = 0.$$

Using that $W_t \rightarrow \phi \hat{a}$ and $\lambda_t \rightarrow \Lambda - \hat{a} - q$ as $t \rightarrow \infty$, we can use (20) to calculate that

$$\hat{F} = \frac{1 - (\gamma - r)\phi \hat{a} - \frac{\phi \hat{a}^2}{2}}{r + \Lambda - \hat{a} - q}, \quad (\text{C.35})$$

which confirms that $\lim_{t \rightarrow \infty} F_t = \hat{F}$. As

$$\hat{a} = \arg \max_{a \in [0, \bar{a}]} \left(\frac{1 - (\gamma - r)\phi a - \frac{\phi a^2}{2}}{r + \Lambda - a - q} \right), \quad (\text{C.36})$$

it follows that optimal effort satisfies $\lim_{t \rightarrow \infty} a_t = \hat{a}$ for an endogenous constant \hat{a} .

Recall the definition of $F^B(q)$ from (B.18). Now note that (C.34) and (C.35) as well as (C.36) jointly imply that $\hat{F} = F^B(q)$ and $\hat{a}^A = a^B(q)$, so that $\hat{W} = W^B(q)$. As a result, it also follows that

$$\lim_{t \rightarrow \infty} V_t = \lim_{t \rightarrow \infty} \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} W_s ds = \frac{\phi \hat{a}}{\gamma + \Lambda - \hat{a} - q} = V^B(q) \quad \text{and} \quad \lim_{t \rightarrow \infty} \dot{V}_t = 0. \quad (\text{C.37})$$

As V_t is the only relevant state variable for the dynamic optimization problem, it follows that V_t cannot have a stationary point $V_t \neq V^B(q)$ with $\dot{V}_t = 0$, as otherwise (C.37) would not hold.

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

Last, we characterize the limit $\lim_{V \rightarrow V^B(q)} F'(V)$. Note that due to (C.25), that is, $F(V^B(q)) = F^B(q)$, and $\lim_{t \rightarrow \infty} V_t = V^B(q)$, it follows that $\lim_{V \rightarrow V^B(q)} F(V) = F^B(q)$ and $\lim_{V \rightarrow V^B(q)} a(V) = a^B(q)$. We know from Proposition 1 that $W^B(q) < F^B(q)$, so that $\lim_{V \rightarrow V^B(q)} W(V) < \lim_{V \rightarrow V^B(q)} F(V)$. Thus, for V close to $V^B(q)$, the principal's limited liability constraint does not bind. Using (C.23), $\lim_{V \rightarrow V^B(q)} a(V) = a^B(q)$ becomes equivalent to

$$\lim_{V \rightarrow V^B(q)} F'(V) = 0, \quad (\text{C.38})$$

when $a^B(q) > 0$. In the case that $a^B(q) = V^B(q) = 0$, we have

$$\lim_{V \rightarrow V^B(q)} F'(V) = \frac{(F^B(q) - (\gamma - r)\phi)}{\phi} \leq 0, \quad (\text{C.39})$$

so that $a(V)$ from (C.23) converges to $a^B(q) = 0$ as $V \rightarrow V^B(q) = 0$.

C.4 Part III

At time $t = 0$, initial screening incentive V_0 pins down screening effort q by means of the screening incentive constraint (10). The agent picks the amount of initial screening incentives V_0 to maximize

$$\max_{q \in [0, \bar{q}]} \left(F(V_0) - \frac{\kappa q^2}{2} \right) \quad \text{s.t.} \quad V_0 = \kappa q. \quad (\text{C.40})$$

Even if optimal screening is not interior and satisfies $q^* = \bar{q}$, it would be optimal to set $V_0 = \kappa q^*$, as $F(V)$ decreases in $V > V^B(q)$ and the screening incentive condition (10) is optimally tight.

The first order condition to (C.40) is

$$\frac{\partial F(V_0)}{\partial q} \Big|_{q=q^*} + F'(V_0)\kappa = \kappa q^*, \quad (\text{C.41})$$

which holds if $q = q^* \in (0, \bar{q})$.

C.5 Part IV

We now explicitly distinguish between q^* (optimal screening level) and q (potentially different screening). This part of the proof shows that in optimum (i.e., for $q = q^*$), we have $\kappa q^* = V_0 > V^B(q^*)$. Because $\lim_{t \rightarrow \infty} V_t = V^B(q^*)$ 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 $t \geq 0$. It suffices to consider $q^* > 0$ and $a^B(q^*) > 0$.

Suppose to the contrary that

$$\kappa q^* = V_0 \leq V^B(q^*) = \frac{W^B(q^*)}{\gamma + \Lambda - a^B(q^*) - q^*}, \quad (\text{C.42})$$

where the last equality follows (C.24). Note that $W_t \leq F_t$ at all times $t \geq 0$ and, in particular, $W^B(q^*) \leq F^B(q^*)$. We then obtain

$$\kappa q^* = V_0 \leq \frac{W^B(q^*)}{\gamma + \Lambda - a^B(q^*) - q^*} < \frac{F^B(q^*)}{r + \Lambda - a^B(q^*) - q^*}, \quad (\text{C.43})$$

where the first inequality follows (C.42) and the second inequality uses $\gamma > r$ and $W^B(q^*) \leq F^B(q^*)$.

Next, define the following (continuous) function (of q):

$$G(q) := F^B(q) - \frac{\kappa q^2}{2}.$$

For any screening effort $q \in (0, \bar{q})$, recall the HJB equation for $V = V^B(q)$, that is, (B.18) or

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

We can use the envelope theorem and differentiate both sides of (B.18) with respect to q to obtain under the optimal controls $(W^B(q), a^B(q))$:

$$(r + \lambda) \frac{\partial F^B(q)}{\partial q} = F^B(q) \iff \frac{\partial F^B(q)}{\partial q} = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} > 0. \quad (\text{C.44})$$

As $a^B(q)$ increases with q (see Proposition 1), above relation implies that $\frac{\partial^2 F^B(q)}{\partial q^2} > 0$ and $\frac{\partial^3 F^B(q)}{\partial q^3} > 0$. Using (C.44), we obtain

$$G'(q) = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q. \quad (\text{C.45})$$

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

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

and satisfies $G'(q^B) = 0$ (first order condition) as well as $G''(q^B) < 0$ (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 $q \in [0, q^B)$.

Next, define the (continuous) function of q :

$$K(q) := V^B(q) - \kappa q, \tag{C.46}$$

with $V^B(q)$ from (C.24), that is,

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

Recall that $a^B(q)$ and $W^B(q) = \phi a^B(q)$ increase with q (see Proposition 1). Thus, the function $V^B(q)$ is strictly convex, implying that $K(q)$ is strictly convex too. Observe that

$$K(q) = V^B(q) - \kappa q = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q} - \kappa q < \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q = G'(q), \tag{C.47}$$

where the first inequality uses that $r < \gamma$ and $W^B(q) \leq F^B(q)$ and the last equality uses (C.45). Because i) $G'(q)$ has a unique root on $[0, q^B]$, ii) because $K(q) < G'(q)$, iii) because $K(q)$ is convex, and iv) because $K(0) \geq 0$, $K(q)$ has a unique root $\hat{q} < q^B$ on $[0, q^B]$ so that $K(\hat{q}) = 0$, $K(q) > 0$ for $q < \hat{q}$, and $K(q) < 0$ for $q \in (\hat{q}, q^B]$. If $K(q)$ had a second root q_2 with $q^B \geq q_2 > \hat{q}$, then it must be due to convexity that $K'(q) > 0$ for $q \geq q_2$ and thus $K(q^B) \geq G'(q^B) = 0$, a contradiction to (C.47).

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

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

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 (15). Because $K(q)$ is strictly convex on $[0, \bar{q}]$, $K(q)$ has precisely one root on $[0, \bar{q}]$, which is denoted \hat{q} and

satisfies $\hat{q} < q^B$. Suppose now $\kappa q^* = V_0 < V^B(q^*)$, which implies $K(q^*) > 0$. Because $K(q)$ has a unique root on $[0, \bar{q}]$, denoted \hat{q} , it follows that $q^* < \hat{q} < q^B$.

Total initial surplus can now be written as

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

where the first inequality uses $F_{0-} \leq F_B(q)$ (which holds for any q) and the second inequality uses that $G(q) = F^B(q) - \frac{\kappa q^2}{2}$ strictly increases on $[0, q^B]$ as well as $0 < q^* < \hat{q} < q^B$. As a result, total surplus is higher under a stationary contract that implements screening \hat{q} and $V_t = V^B(\hat{q}) = \kappa \hat{q}$ 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 $q^* = \hat{q} < q^B$. Take $\varepsilon > 0$ and set $q^\varepsilon = q^* + \varepsilon$ so that $q^\varepsilon < q^B$. Because of $q^* < q^B$, it follows that

$$\frac{\partial}{\partial q^*} \left(F^B(q^*) - \frac{\kappa(q^*)^2}{2} \right) = G'(q^*) > 0, \quad (\text{C.48})$$

where $G(q^*) = F^B(q^*) - \frac{\kappa(q^*)^2}{2}$ 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

$$\begin{aligned} F(V_0) - \frac{\kappa(q^\varepsilon)^2}{2} &= F^B(q^\varepsilon) + F'(V^B(q^\varepsilon))\varepsilon + o(\varepsilon^2) - \frac{\kappa(q^\varepsilon)^2}{2} = F^B(q^\varepsilon) + o(\varepsilon^2) - \frac{\kappa(q^\varepsilon)^2}{2}, \\ &= \left(F^B(q^*) - \frac{\kappa(q^*)^2}{2} \right) + \frac{\partial}{\partial q^*} \left(F^B(q^*) - \frac{\kappa(q^*)^2}{2} \right) \varepsilon + o(\varepsilon^2), \end{aligned} \quad (\text{C.49})$$

which — by (C.48) — exceeds $F^B(q^*) - \frac{\kappa(q^*)^2}{2}$ for $\varepsilon > 0$ sufficiently small. The second equality uses that given screening level q^ε , $\lim_{V \rightarrow V^B(q^\varepsilon)} F'(V) = 0$ (see (C.38)) which holds because of $a^B(q^\varepsilon) > 0$ which in turn follows from $a^B(q^*) > 0$ by continuity for small ε . 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.6 Part V

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

First, consider $F(V) = W(V)$, in that the principal's limited liability constraint binds. The expression for effort $a(V) = W(V)/\phi$ in (C.23) implies that $F'(V) < 0$, because $F'(V) \geq 0$ would imply $a(V) < F(V)/\phi$ and $W(V) < F(V)$. Next, take $F(V) = W(V) = \phi a(V)$ and insert this relation into the HJB equation (22) to obtain

$$\gamma F(V) = 1 - \frac{F(V)^2}{2\phi} - \left(\Lambda - q - \frac{F(V)}{\phi} \right) F(V) + F'(V) \left[\left(\gamma + \Lambda - q - \frac{F(V)}{\phi} \right) V - F(V) \right].$$

At points V at which $F'(V)$ is differentiable, we can differentiate above ODE with respect to V to

calculate

$$F''(V) = \frac{(F'(V))^2 - F'(V)F(V)/\phi + (F'(V))^2V/\phi}{(\gamma + \lambda)V - F(V)} < 0,$$

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

Second, suppose that $F(V) > W(V)$ and the principal's limited liability constraint does not bind, and consider $V > V^B(q)$. To start with, note that because the principal's limited liability constraint does not bind, optimal effort $a(V)$ solves the first order condition $\frac{\partial F(V)}{\partial a} = 0$ provided $a \in (0, \bar{a})$. For any points V at which $F'(V)$ is differentiable, we can then invoke the envelope theorem and totally differentiate the HJB equation (22) under the optimal controls with respect to V , which yields

$$F''(V) = \frac{-(\gamma - r)F'(V)}{(\gamma + \lambda)V - W}. \quad (\text{C.50})$$

First, note that as shown in Part II of the proof, $\dot{V} = (\gamma + \lambda)V - W < 0$ for $V > V^B(q)$. Thus, $F''(V)$ has the same sign as $F'(V)$. It follows by (C.50) that either $F'(V), F''(V) < 0$ or $F'(V), F''(V) \geq 0$ must hold for all $V \in (V^B(q), V_0]$.

Next, let us consider $V = V^B(q)$ (or the limit $V \rightarrow V^B(q)$). When $a^B(q) = 0$, then (C.39) implies $\lim_{V \rightarrow V^B(q)} F'(V) \leq 0$. Otherwise, when $a^B(q) > 0$, the N (C.38) implies $F'(V^B(q)) = 0$ and — according to the expression for effort (C.23):

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

owing to $\gamma > r$.

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

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

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

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

C.7 Part VI

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

We can solve (12) to get the payout rate

$$c_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t. \quad (\text{C.51})$$

If $F_t = W_t$, note that according to (B.14), $\dot{F}_t = (\gamma + \lambda_t)F_t - 1 + \frac{\phi a_t^2}{2}$. Inserting the law of motion $\dot{F}_t = \dot{W}_t$ into (C.51) yields $c_t = 1 > 0$.

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

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

and, provided $a(V)$ is differentiable, then $a'(V) = \frac{-F''(V)[V + \phi]}{\phi} > 0$, as $F''(V) < 0$ when $W < F(V)$. Thus, $\dot{a}_t = a'(V_t)\dot{V}_t < 0$ and, by (7), $\dot{W}_t < 0$. Inserting $\dot{W}_t < 0$ into (C.51) implies $c_t > 0$.

D Additional results

D.1 Proof of Corollary 1

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

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

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

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

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

Next, note that

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

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

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

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

solves the ODE

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

subject to the boundary condition

$$\lim_{V \rightarrow V^B(q)} L'(V) = 0 \iff \lim_{V \rightarrow V^B(q)} L(V) = \frac{1}{r + \Lambda - a^B(q) - q}.$$

Second, calculate

$$\dot{W}_t = W'(V_t)\dot{V}_t \quad \text{and} \quad \dot{\beta}(V) = \beta'(V_t)\dot{V}_t,$$

where $\beta(V)$ is the agent's retention level in state V under the proposed implementation of the optimal contract. Third, insert these relations into (29) to obtain the following ODE in state V

$$\beta(V) - \beta'(V)\dot{V}L(V) = (\gamma + \Lambda - a(V) - q)W(V) + \frac{\phi a(V)^2}{2} - W'(V)\dot{V}, \quad (\text{D.52})$$

which is solved subject to

$$\lim_{V \rightarrow V^B(q)} \beta(V) = 0 \iff \lim_{V \rightarrow V^B(q)} \beta(V) = c^B(q) = (\gamma + \Lambda - a^B(q) - q)W^B(q) + \frac{\phi(a^B(q))^2}{2}. \quad (\text{D.53})$$

Noting there is a one-to-one mapping from time t to $V_t = V$, we thus obtain $\beta_t = \beta(V_t)$ by solving (D.52), as desired. Under standard regularity conditions, well-known results imply the existence of a solution of the ODE (D.52) subject to (D.53); throughout, we assume the existence and uniqueness of such a solution.

D.3 Model variant with only moral hazard over screening

D.3.1 Solution

We characterize the model solution when there is no moral hazard over monitoring (i.e., monitoring effort a_t is contractible), so that the incentive constraint (7) does not apply. However, there is still moral hazard over screening, i.e., q is unobserved and not contractible. 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.

The agent's continuation payoff follows²⁰

$$dW_t = (\gamma + \lambda_t)W_t dt + \frac{\phi a_t^2}{2} dt - dC_t, \quad (\text{D.54})$$

with payouts dC_t . Noting that an unobserved change in q does not affect contracted monitoring effort a_t (i.e., $\frac{\partial a_t}{\partial q} = \frac{\partial dC_t}{\partial q} = 0$), we can differentiate this law of motion (D.54) with respect to screening effort q to obtain (after simplifications) for $V_t = \frac{\partial}{\partial q} W_t$:

$$\dot{V}_t = (\gamma + \lambda_t)V_t - W_t,$$

²⁰Since both dC_t and a_t are contractible, one could define $d\hat{C}_t := dC_t - \frac{\phi a_t^2}{2} dt$ and write $dW_t = (\gamma + \lambda_t)W_t dt - d\hat{C}_t$, where $d\hat{C}_t$ is a (contracted) choice variable.

which is dynamics of the agent's screening incentives. At time $t = 0$, the incentive constraint $V_0 = \kappa q$ pins down screening effort.

As in the baseline, the agent maximizes total surplus at time $t = 0$. The only relevant state variable is V , while W is control variable. As such, total surplus (i.e., the value function) is a function of V only and solves the HJB equation

$$rF(V) = \max_{W \in [0, F(V)], a \in [0, \bar{a}]} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)((\gamma + \lambda)V - W) \right\}, \quad (\text{D.55})$$

which is analogous to the baseline HJB equation (22). The key difference to the baseline (where the incentive condition $W = \phi a$ links monitoring effort and continuation value) is that without moral hazard over monitoring (i.e., with contractible a) the monitoring incentive constraint does not apply and W and a can be chosen independently in the optimization in (D.55). In what follows, we assume that a unique solution to (D.55) (subject to a boundary condition specified later) exists.

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

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

Note that (D.55) implies

$$\frac{\partial rF(V)}{\partial W} = -(\gamma - r) + F'(V).$$

As such, the maximization with respect to the agent's deferred compensation, i.e., W , in (D.55) yields that

$$W(V) \begin{cases} = 0 & \text{if } F'(V) > -(\gamma - r) \\ \in [0, F(V)] & \text{if } F'(V) = -(\gamma - r) \\ = F(V) & \text{if } F'(V) < -(\gamma - r). \end{cases} \quad (\text{D.56})$$

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 $\lim_{t \rightarrow \infty} V_t = V^B(q) = 0$, i.e., given q , the optimal contract approaches in the limit $t \rightarrow \infty$ the one with contractible screening. 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 (D.56), this requires that $W(V) > 0$ for $V > 0$ and so $F'(V) \leq -(\gamma - r)$ for $V > 0$.

Thus, it is (at least) weakly optimal to stipulate $W(V) = F(V)$, which we can insert into the HJB equation (D.55) to obtain

$$\gamma F(V) = \max_{a \in [0, \bar{a}]} \left\{ 1 - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)((\gamma + \lambda)V - F(V)) \right\}. \quad (\text{D.57})$$

Let us assume that $F''(V)$ exists and is well-defined. Using the envelope theorem, we totally differentiate the HJB equation (D.57) (under the optimal control $a = a(V)$) with respect to V ,

which yields

$$F''(V) = \frac{(F'(V))^2}{(\gamma + \lambda)V - F(V)}.$$

Due to $\dot{V} = (\gamma + \lambda)V - F(V) < 0$, we have $F''(V) < 0$, i.e., $F(V)$ is strictly concave. That is, $F(V)$ is strictly concave for $V > 0$. If there exists now $\hat{V} > 0$ 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 $V > 0$, so that — indeed — $W(V) = F(V)$ is optimal for $V > 0$.

When V equals zero, it must be that \dot{V} equals zero too, as — by definition — V cannot become negative. As such, $W(0) = 0$, which requires by means of (D.56) that $F'(0) \geq -(\gamma - r)$. As $F'(V) < -(\gamma - r)$ and $F'(V)$ is continuous for all $V > 0$, it follows that $F'(0) = -(\gamma - r)$ which is the boundary condition for the ODE (D.55). Notice that this boundary condition is equivalent to

$$\lim_{V \rightarrow 0} F(V) = \max_{a \in [0, \bar{a}]} \left(\frac{1 - \frac{\phi a^2}{2}}{r + \Lambda - a - q} \right), \quad (\text{D.58})$$

which—given the level of q —is total surplus absent any moral hazard. Also observe that because $W(V) = F(V) > W(0)$ for $V > 0$ with $\lim_{V \downarrow 0} W(V) > 0$, it follows that $\lim_{V \downarrow 0} \dot{V}(V) > 0 = \dot{V}(0)$; thus, state $V = 0$ is reached in finite $\tau^0 = \inf\{t \geq 0 : V_t = 0\}$.

Finally, we can determine optimal q . 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 $V_0 = \kappa q$.

D.3.2 Implementation of the optimal contract

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

Next, we characterize the payouts to the agent and, doing so, we omit time subscripts unless confusion is likely to arise. Recall from the previous section that

$$F(0) = \lim_{V \downarrow 0} F(V) = \lim_{V \downarrow 0} W(V) > W(0) = 0.$$

Using the law of motion for the agent's continuation payoff

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

it follows that the agent receives a payout $dC = F(0)$ at the time V reaches zero, so as to induce $F(0) = \lim_{V \downarrow 0} W(V) > W(0) = 0$. When $V > 0$, then $F(V) = W(V)$, and according to (B.14) for $W(V) = F(V)$:

$$dW = (\gamma + \lambda)W dt + \frac{\phi a^2}{2} dt - dC = (\gamma + \lambda)F dt + \frac{\phi a(V)^2}{2} dt - 1 dt = dF,$$

yielding

$$dC = 1 dt,$$

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 $\tau^0 = \inf\{t \geq 0 : V_t = 0\}$ and to sell them to outside investors at the time V reaches zero. When $V = 0$ at time τ^0 , 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.4 Proof of Proposition 4

The first claim follows from Proposition 1: It readily follows that the optimal contract can be implemented by having the agent retain constant share $\beta_t = c^B(q)$ of the loan. The second claim follows from the solution as well as the implementation of the optimal when there is no moral hazard over monitoring, presented in Appendix D.3.

D.5 Model extension with finite maturity

D.5.1 Solution

We now provide additional details, the solution, and derivations for the model variant with finite debt maturity where $\delta > 0$. The incentive constraints with respect to monitoring and screening effort remain unchanged relative to the baseline, i.e., $W_t = \phi a_t$ and $V_0 = \kappa q$, pinning down $\lambda_t = \Lambda - a_t - q$. To solve the model, one first takes q as given to characterize the solution after time $t = 0$; then, taking into account the continuation solution, one maximizes initial surplus $F_{0-} = F_0 - \frac{\kappa q^2}{2}$ over q .

To begin with, we define the agent's continuation value (before maturity) as

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

where dC_s^δ is the agent's payoff in the form of a lump-sum payment upon maturity (which occurs randomly at rate δ) at time s and c_s the payout rate before maturity (we conjecture and verify that payments before maturity are smooth). Observe that over $[t, t + dt)$, the loan matures with probability δdt in which case the agent is paid dC_t^δ dollars (note that dC_t^δ is not of order dt).

Differentiating above expression with respect to time, t , we obtain:

$$\dot{W}_t = (\gamma + \delta + \lambda)W_t + \frac{\phi a_t^2}{2} - c_t - \delta dC_t^\delta. \quad (\text{D.59})$$

According to the dynamic programming principle, the agent solves at any time t the optimization:

$$(\gamma + \delta)W_t = \max_{a_t \in [0, \bar{a}]} \left(c_t - \lambda_t W_t - \frac{\phi a_t^2}{2} + \delta dC_t^\delta + \dot{W}_t \right), \quad (\text{D.60})$$

yielding $a_t = W_t/\phi$ (if monitoring effort is interior).

Note also that because screening effort q is neither observable nor contractible, an unobserved change in screening effort q cannot affect contracted flow payments c_t or the lump-sum payment dC_t^δ upon maturity. Using the envelope theorem (i.e., $\frac{\partial}{\partial q} \frac{\partial W_t}{\partial a_t} = 0$) and $\frac{\partial c_t}{\partial q} = \frac{\partial dC_t^\delta}{\partial q} = 0$, we can

differentiate both sides of above equation (D.60) with respect to q to obtain for $V_t = \frac{\partial}{\partial q} W_t$:²¹

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

which is (32). Equivalently, we obtain the integral representation

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

which becomes (31) for $t = 0$.

Next, we denote the continuation surplus after maturity at a time s by F_s^δ . Thus, the continuation surplus at time t before maturity is characterized in (30), i.e.,

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

By the dynamic programming principle, the value function $F_t = F(V_t, W_t)$ solves the HJB equation

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

As in the baseline, the optimality of payouts requires

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

Recall that ex-ante, we do not restrict c to be positive, but afterward verify that $c \geq 0$.

With slight abuse of notation, we write $F_t = F(V_t)$ (i.e., F_t is a function of V_t only) and using $F_W = 0$, the HJB equation simplifies to

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

with $W = \phi a$ and $W \leq F(V)$ (limited liability).

²¹An alternative derivation (not relying explicitly on envelope theorem) simply rewrites (D.59) by inserting monitoring incentive compatibility, $a_t = W_t/\phi$, to obtain

$$\dot{W}_t = \left(\gamma + \delta + \Lambda - \frac{W_t}{\phi} - q \right) W_t + \frac{W_t^2}{2\phi} - c_t - \delta dC_t^\delta.$$

Differentiating both sides with respect to q and using $\frac{\partial c_t}{\partial q} = \frac{\partial dC_t^\delta}{\partial q} = 0$, we obtain

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

As in the baseline, the state variable V_t converges to a limit $V^B(q)$, i.e., $\lim_{t \rightarrow \infty} V_t = V^B(q)$, whereby $\lim_{t \rightarrow \infty} \dot{V}_t = 0$.²² Then, the HJB equation (D.61) is subject to the boundary condition

$$\lim_{V \rightarrow V^B(q)} F(V) = F^B(q) = \max_{W \in [0, F^B(q)]} \left(\frac{1 + \delta F^\delta}{r + \Lambda - a - q + \delta} - \frac{(\gamma - r)W}{r + \Lambda - a - q + \delta} - \frac{\frac{\phi a^2}{2}}{r + \Lambda - a - q + \delta} \right), \quad (\text{D.62})$$

which is analogous to (23) in the baseline model. Here,

$$V^B(q) = \frac{W^B(q)}{r + \delta + \Lambda - a^B - q} \quad \text{with} \quad W^B(q) = W(V^B(q)) \quad \text{and} \quad a^B(q) = \frac{W^B(q)}{\phi}. \quad (\text{D.63})$$

We assume that a unique solution to (D.61) (subject to above boundary condition) exists.

In addition, as in the baseline model, optimal screening effort $q^* = q$ maximizes total initial surplus $F_{0-} = F(V_0) - \frac{\kappa q^2}{2}$ subject to the incentive constraint $V_0 = \kappa q$. We numerically verify that (under the chosen parameters) in optimum, $V_0 \geq V^B(q)$, so that $\dot{V}_t < 0$ and V_t drifts down over time $V^B(q)$, as well as that the value function is strictly concave and decreases (i.e., $F'(V), F''(V) < 0$). A rigorous proof could be constructed using analogous arguments as those presented in the proof of Proposition 2.

In what follows, we assume for simplicity that $F_s^\delta = F_s$ (or $F^\delta = F(V)$), i.e., the stochastic maturity event leaves the total loan value unchanged, in which case (20) and (30) coincide. At maturity, the lender is paid W_t and outside investors are paid $F_t - W_t$. Therefore, there is no value effect associated with the maturity event.²³ This assumption reflects in reduced form the fact that the value of the loan is the same just before maturity and at maturity; in a model with a deterministic maturity date, this property would be called a value matching condition.²⁴

Thus, using $F^\delta = F(V)$, the HJB equation (D.61) simplifies to

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

with $W = \phi a$ and $W \leq F$ (limited liability). The boundary condition (D.62) simplifies to

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

Optimal effort becomes

$$a(V) = \frac{F(V) - F'(V)(V + \phi) - (\gamma - r)\phi}{\phi} \quad \wedge \quad W(C).$$

²²We numerically verify that, indeed, $\dot{V}_t < 0$. A formal proof could be constructed using arguments analogous to those in the proof of Proposition 2.

²³This assumption has no bearings on our key findings and is for mere simplicity; our results would remain qualitatively unchanged had we assumed different F_t^δ , for instance, $F_t^\delta = K$ for a constant $K \geq 0$.

²⁴In reality, loans mature deterministically and this feature naturally holds, preventing arbitrage.

It follows that $a'(V) \geq 0$ as well as $\dot{a}, \dot{W} < 0$. The exact level of dC_t^δ (or dC^δ) is payoff-irrelevant and does not affect key equilibrium quantities, such as total surplus, credit risk, and screening or monitoring incentives. Thus, we can without loss of generality set $dC^\delta = W$, i.e., we assume that the maturity event does not change the agent's continuation value just as it does not change the value of total surplus due to $F^\delta = F(V)$.

Finally, we can calculate the retention level β_t via

$$\beta_t - \dot{\beta}_t L_t = c_t \iff \beta(V) - c(V) = L(V)\beta'(V)\dot{V},$$

where the market value of debt, $L_t = L(V_t)$, is defined as

$$L_t = \int_t^\infty e^{-(r+\delta)(s-t) - \int_t^s \lambda_u du} (1 + \delta L_s^\delta) ds$$

and payouts to the agent, $c_t = c(V_t)$, read (after using $dC_t^\delta = W_t$ in (D.59))

$$c_t = (\gamma + \lambda)W + \frac{\phi a^2}{2} - \dot{W} \geq 0.$$

Here, L_s^δ is the market value of debt at the maturity event (i.e., the “face value” repaid to lenders at maturity). For simplicity, we assume — in line with $F_s^\delta = F_s$ and $dC_s^\delta = W_s$ — that $L_s^\delta = L_s$, leading to $L_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} 1 ds$. That is, the maturity event is value neutral for total surplus $F(V)$, agent continuation value $W(V)$, and the value of debt.

D.5.2 Main results and figures with finite maturity

We now replicate Figures 2 and 4 for finite maturity, where we choose $\delta = 0.1$. Similar to Figure 2 in the baseline (infinite maturity) case, Figure D.1 plots screening and monitoring effort against κ , ϕ , Λ , and γ . Indeed, as Figure D.1 illustrates, monitoring and screening efforts decrease with κ , ϕ , Λ , and γ , producing qualitatively similar patterns as Figure 2 does.

Next, similar to Figure 4 in the baseline (infinite maturity) case, Figure D.2 plots retention levels and selloff speed against κ , ϕ , Λ , and γ . Again, it can be seen that Figure D.2 produces qualitatively similar results as Figure 4 does. As such, we conclude that our model's key results (on effort incentives and retention dynamics) are robust to the level of loan maturity.

D.6 Model variant with separation of screening and monitoring

We now assume that screening and monitoring are undertaken by two separate agents, referred to as the screener and monitor respectively. Both the screener and monitor have identical preferences, i.e., they are risk-neutral with discount rate γ . Both screening q and monitoring a_t are not observable nor contractible, and affect default rate $\lambda_t = \Lambda - a_t - q$. That is, only the screener (monitor) observes screening (monitoring) effort q (a_t). A contract to the screener \mathcal{C}^s stipulates recommended screening \hat{q} and incremental payouts dC_t^s ; a contract to the monitor \mathcal{C}^m stipulates recommended monitoring $\{\hat{a}_t\}$ and incremental payouts dC_t^m . The contracts are chosen to maximize total surplus. We focus on incentive compatible contracts, so that in optimum $q = \hat{q}$ and $a_t = \hat{a}_t$.

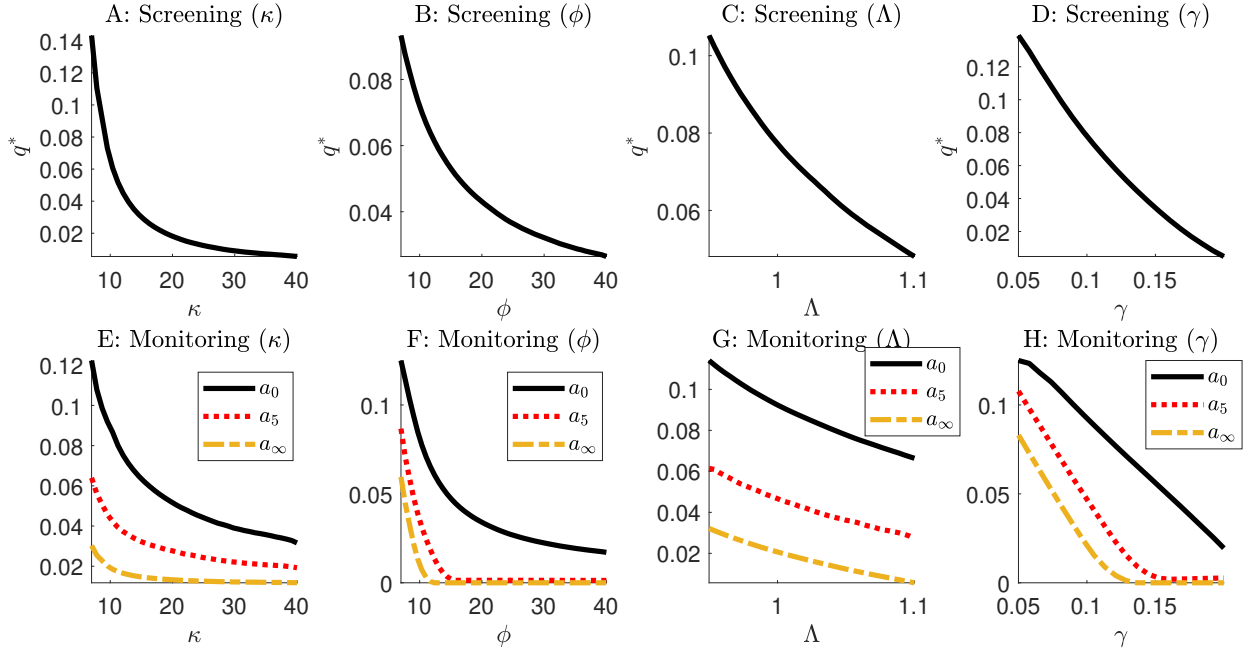


Figure D.1: **Comparative statics with finite maturity.** This figure plots monitoring effort a_t at $t = 0$ (solid black line), at $t = 5$ (dotted red line), and $t \rightarrow \infty$ (dashed yellow line) and screening effort q^* against the parameters ϕ, κ, Λ , and γ . We use our baseline parameters and set $\delta = 0.1$.

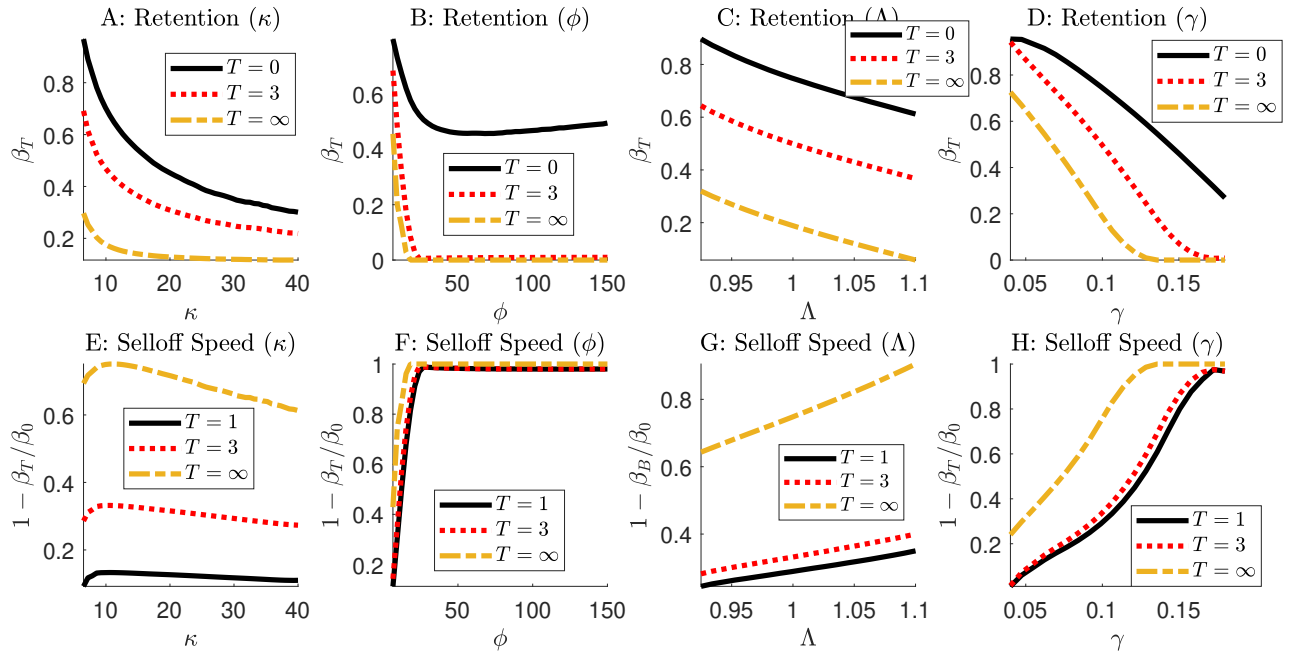


Figure D.2: **Retention and dynamics with finite maturity.** We use our baseline parameters and $\delta = 0.1$.

D.6.1 Optimals contract and solution with separation of screening and monitoring

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 assume that monitoring effort (screening effort) is only and privately observed by the monitor (screener).

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

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

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

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

where a_t is monitoring effort and q is screening effort, leading to $\lambda_t = \Lambda - a_t - q$. The loan matures randomly at rate δ , and $dC_t^{s,\delta}$ and $dC_t^{m,\delta}$ are the screener's and monitor's payoffs (lump-sum payments) in the event of maturity respectively (note that $dC_t^{s,\delta}$ and $dC_t^{m,\delta}$ are not of order dt). That is, over $[t, t + dt)$, the loan matures with probability δdt in which case the screener (monitor) is paid $dC_t^{s,\delta}$ ($dC_t^{m,\delta}$) dollars.

As such, we obtain the following dynamics for continuation values:

$$dW_t^s = (\gamma + \lambda_t + \delta)W_t^s dt - dC_t^s - \delta dC_t^{s,\delta} dt \quad (\text{D.64})$$

$$dW_t^m = (\gamma + \lambda_t + \delta)W_t^m dt - dC_t^m + \frac{\phi a_t^2}{2} dt - \delta dC_t^{m,\delta} dt. \quad (\text{D.65})$$

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 . Moreover, as will become clear later, the exact values of the payments $\delta dC_t^{s,\delta}$ and $\delta dC_t^{m,\delta}$ will turn out not to be relevant for key equilibrium quantities, such as incentives, credit risk, or total surplus.

At any point in time, the monitor chooses effort a_t to maximize

$$(\gamma + \delta)W_t^m = \max_{a_t \in [0, \bar{a}]} \left(\lambda_t W_t^m + dC_t^m + \delta dC_t^{m,\delta} + \frac{dW_t^m}{dt} \right).$$

Thus, optimal monitoring (if interior) is pinned down by the incentive condition

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

provided that monitoring effort a_t is interior. Next, the screener maximizes at time $t = 0$:

$$\max_{q \in [0, \bar{q}]} W_0 - \frac{\kappa q^2}{2},$$

As in the baseline version of the model, optimal screening is pinned down by the incentive condition

$$V_0 = \kappa q,$$

where we define $V_t := \frac{\partial}{\partial q} W_t^s$ as the screener's "screening" incentives. The remainder of the solution, similar to the baseline, features V_t as the main state variable, and W_t^s and W_t^m are control variables in the dynamic optimization.

Noting that an unobserved change in screening effort does not affect contracted payments, so that $\frac{\partial dC_t^s}{\partial q} = \frac{\partial dC_t^{s,\delta}}{\partial q} = 0$, or the monitor's monitoring effort, so that $\frac{\partial a_t}{\partial q} = 0$, we can differentiate the dynamics of W_t^s in (D.64) with respect to q to obtain for the screener's incentives $V_t := \frac{\partial W_t^s}{\partial q}$:

$$dV_t = (\gamma + \lambda_t + \delta)V_t dt - W_t^s dt. \quad (\text{D.66})$$

Thus, the screener's "screening" incentives in integral form read

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

The optimal contracts to both the screener and monitor are designed to dynamically maximize total surplus F_t . Total surplus F_t can be rewritten (using arguments analogous to the ones that lead to (B.16)) as

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

where F_s^δ is the (continuation) surplus "just after" maturity (which occurs at rate δ). We will specify the exact form of F_s^δ below.

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

$$(r + \delta)F(V) = \max_{a, W^m, W^s} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)(W^m + W^s) - \lambda F(V) + \delta F^\delta + F'(V)((\gamma + \lambda + \delta)V - W^s) \right\}. \quad (\text{D.67})$$

Note that limited liability requires that $W^m \in [0, F(V) - W^s]$ and $W^s \in [0, F(V) - W^m]$ and incentive compatibility with respect to monitoring requires that $W^m = a\phi$. Throughout, we assume existence and uniqueness of a solution to (D.67) (subject to a boundary condition specified below).

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

$$W^s(V) \begin{cases} = 0 & \text{if } F'(V) > -(\gamma - r) \\ \in [0, F(V) - W^m(V)] & \text{if } F'(V) = -(\gamma - r) \\ = F(V) - W^m(V) & \text{if } F'(V) < -(\gamma - r). \end{cases} \quad (\text{D.68})$$

As in the baseline, it follows that $\lim_{t \rightarrow \infty} V_t = V^B(q)$, where $V^B(q)$ is the level of screening

incentives in the benchmark without screening moral hazard (given q).²⁵ 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 $t \geq 0$.

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

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

Owing to (D.68), this requires that $W^s(V) > 0$ for $V > 0$ and therefore $F'(V) \leq -(\gamma - r)$ for $V > 0$. Next, suppose that $F'(V) < -(\gamma - r)$ for $V > 0$, so $W^s(V) = F(V) - W^m(V)$. Inserting this expression into (D.67) and simplifying leads to the ordinary differential equation

$$(\gamma + \delta)F(V) = \max_{a, W^m} \left\{ 1 - \frac{\phi a^2}{2} - \lambda F(V) + \delta F^\delta + F'(V)((\gamma + \lambda + \delta)V - F(V) + W^m) \right\}, \quad (\text{D.69})$$

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

As in the main text (compare Section 4), we consider $F^\delta = F(V)$, so (D.69) simplifies to

$$\gamma F(V) = \max_{a, W^m} \left\{ 1 - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)((\gamma + \lambda + \delta)V - F(V) + W^m) \right\}. \quad (\text{D.70})$$

Using the envelope theorem to totally differentiate the HJB equation (D.70) (under the optimal control $W^m = \phi a$) with respect to V yields

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

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 (D.66)). 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 (D.69) with respect to monitoring effort yields

$$a(V) = \frac{F(V) - F'(V)V + F'(V)\phi}{\phi}. \quad (\text{D.71})$$

When V approaches zero, it must be that \dot{V} approaches zero too, as — by definition — V cannot become negative. As such, $W^s(0)$ approaches zero, which requires by means of (D.68) that $F'(0) \geq -(\gamma - r)$. As $F'(V) < -(\gamma - r)$ for all $V > 0$, it follows — by continuity of $F'(V)$ — that $\lim_{V \rightarrow 0} F'(V) = -(\gamma - r)$. An alternative way to derive this boundary condition is as follows. Comparing (18) with (D.69), one can see that

$$\lim_{V \rightarrow 0} F(V) = F^B(q) = \max_{a \in [0, \bar{a}]} \left(\frac{1 - (\gamma - r)\phi a - 0.5\phi a^2}{r + \Lambda - a - q} \right)$$

²⁵We omit the formal proof of this claim which could be constructed using arguments analogous to those presented in Part II of Proposition 2.

is equivalent to

$$\lim_{V \rightarrow 0} F'(V) = -(\gamma - r),$$

which is then natural the boundary condition for the ODE (D.69) as V approaches zero. We assume that a unique solution to (D.69) (subject to above boundary condition) exists.

Finally, notice that the exact values of the payoffs upon maturity, i.e., $dC_t^{m,\delta}$ and $dC_t^{s,\delta}$, are not payoff-relevant, in a sense that they do not affect monitoring or screening incentives, credit risk, or total surplus. Thus, as in Appendix D.5, we can assume that the maturity event does not change the agent payoff, i.e., we stipulate $dC_t^{s,\delta} = W_t^s$ and $W_t^{m,\delta} = W_t^m$. Again, this assumption is without loss of generality, since the exact values of $dC_t^{s,\delta}$ and $W_t^{m,\delta}$ do not affect key equilibrium quantities, such as total surplus, credit risk, and screening or monitoring incentives.

The screener's continuation payoff follows then the dynamics

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

Because $\lim_{V \downarrow 0} W^s(V) \geq W^s(0) = 0$, the the screener receives a payout of

$$dC^s = W^s(0) = F(0) - W^m(0)$$

dollars at the time V reaches zero, which occurs in finite time owing to $\lim_{V \downarrow 0} \dot{V}(V) > 0 = \dot{V}(0)$.

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 $V_0 = \kappa q$.

D.6.2 Contract dynamics with separation of screening and monitoring

We show that when screening and monitoring are separate and $\phi > \kappa \bar{q}$, then monitoring effort increases over time, i.e., $a'(V) < 0$ and $\dot{a}_t > 0$, so that credit and default risk decrease over time, as opposed to the baseline in which monitoring effort decreases and credit risk increases over time.

Recall the monitoring effort from (D.71), that is,

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

We can differentiate $a(V)$ with respect to V to obtain

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

As $q < \bar{q}$ and $\dot{V}_t \leq 0$, we have $V_t < V_0 \leq \kappa \bar{q}$. Moreover, the value function is strictly concave, i.e., $F''(V) < 0$, and — by assumption — $\phi > \kappa \bar{q}$ holds, so that

$$a'(V) \leq \frac{-F''(V)(\kappa \bar{q} - \phi)}{\phi} < 0.$$

Thus, effort a_t increases over time, i.e., $\dot{a}_t = a'(V_t)\dot{V}_t > 0$.