The Dynamics of Loan Sales and Lender Incentives*

Sebastian Gryglewicz† Simon Mayer‡ Erwan Morellec§

October 3, 2023

Abstract

How much of a loan should a lender retain and how do loan sales 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. The optimal lender-investor contract can be implemented by requiring the lender to initially retain a share of the loan that it gradually sells to investors, rationalizing loan sales after origination. The model generates novel predictions linking loan and lender characteristics to initial retention, sales dynamics, and loan performance.

Keywords: dynamic agency, screening, monitoring, loan sales, syndicated loans.

JEL Classification: G21, G32.

*We would like to thank Itay Goldstein (the editor), two anonymous referees, Bo Becker, Bruno Biais, Matthieu Bouvard, Will Cong, Doug Diamond, Matthias Efing, Quirin Fleckenstein, Andreas Fuster, Thomas Geelen, Denis Gromb, Barney Hartman-Glaser, Alexander Guembel, Kinda Hachem, Sharjil Haque, Florian Hoffmann, Shohini Kundu, Gustavo Manso, Andrey Malenko, Nadya Malenko, Ralf Meisenzahl, Maureen O'Hara, Martin Oehmke, Cecilia Parlatore, Amiyatosh Purnanandam, Uday Rajan, Alejandro Rivera, Anthony Saunders, Philip Schnabl, Sascha Steffen, Per Stroemberg, Stephane Villeneuve, Vish Viswanathan, Mao Ye, and seminar participants at Carnegie Mellon University (Tepper), Cornell University (SC Johnson College of Business), HEC Paris, the University of Bonn, the University of Michigan (Ross School of Business), the University of Rochester (Simon School of Business), Stockholm School of Economics, Toulouse School of Economics, MFA 2022, and the 2022 FTG meeting in Budapest for comments. Erwan Morellec acknowledges financial support from the Swiss Finance Institute. Part of this research has been completed while Erwan Morellec was a visiting professor of finance at the MIT Sloan School of Management. The paper was previously circulated under the title "Screening and monitoring corporate loans."

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

‡Tepper School of Business, Carnegie Mellon University. E-mail: mayer@hec.fr.

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

Banks provide unique services in the form of publicly unobservable screening and monitoring of borrowers. A central result in banking theory is that for banks to have the incentive to provide an efficient level of these services, it is necessary for them to retain part of the loans they originate (Gorton and Pennacchi, 1995). Lenders who sell loans to investors will bear fewer costs in the event of default and therefore may have less incentive to screen or monitor borrowers.

The view that banks have significant skin in the game and therefore provide an efficient level of these services has been challenged by recent developments in the market for corporate loans. Indeed, the emergence of an active and liquid secondary market for corporate loans (Saunders, Spina, Steffen, and Streitz, 2021) has given banks the possibility to reduce their exposure to borrowers' default risk by selling their stake over the loan's life (Drucker and Puri, 2009; Nadauld and Weisbach, 2012; Irani, Iyer, Meisenzahl, and Peydro, 2021). As further shown by Blickle, Fleckenstein, Hillenbrand, and Saunders (2022), in the syndicated loan market,[1] lead banks sell their entire share shortly after origination for a significant fraction of the loans they syndicate. Several important questions naturally arise in this context. First, what determines optimal initial retention for loan originators, as well as retention dynamics and loan sales after origination? Second, how do loan sales affect moral hazard in screening and monitoring, and therefore loan performance and value?

This paper attempts to answer these questions by developing a tractable, unifying framework of loan origination and sales under moral hazard in screening and monitoring. Our model applies to corporate loans, and in particular to syndicated corporate loans, but is sufficiently general to apply to other credit markets, such as mortgage loans and their securitization.[2] We then use this framework to characterize the dynamically optimal originator

[1] Syndicated 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 (see, e.g. Sufi (2007) or Saunders et al. (2021)).

[2] As documented for instance in Benmelech, Dlugosz, and Ivashina (2012), the securitization of corporate loans—most commonly structured as collateralized loan obligations (CLOs)—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. The fractions of the same underlying loan are held simultaneously by CLOs as well as by

share and its relation to moral hazard and loan performance. This allows us to (*i*) shed light on recent empirical findings and (*ii*) generate new predictions regarding optimal retention by loan originators, the dynamics of loan sales by originators, their relation to loan characteristics, and their effects on loan performance and value.

We start our analysis by formulating a dynamic agency model in which a lender—the lead bank in a loan syndicate—originates a loan and sells this loan to competitive investors—other banks in the syndicate or non-bank financial intermediaries. The loan generates coupon payments at a constant rate until default or maturity. The lender may undertake a costly screening effort at origination that results in a lower expected default rate at all future times. 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 screening and monitoring efforts. Because screening and monitoring are not observable, there is moral hazard and the lender's 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 (part of) the loan to investors. However, loan sales reduce the lender's exposure to loan performance and undermine its incentives to screen and monitor, thereby increasing credit risk and reducing loan value.

We derive the optimal contract between the lender (loan originator) and outside investors that implements costly screening and monitoring, while respecting the limited liability of the lender and investors. Incentive provision requires exposing the lender to loan performance. As the lender is protected by limited liability, this is achieved by delaying its payouts so that the lender loses its expected future payouts upon default. However, delaying payouts 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. Under the optimal contract, the selloff speed decreases over time, so most loan sales occur

other institutional investors and banks. Furthermore, each loan included in CLOs is rated.

relatively shortly after origination, in line with observed practice.

The structure of the optimal contract reflects the fact that screening only occurs at origination, so that the contract front-loads incentives. Therefore, the lender's exposure to loan performance and incentives to monitor are especially strong at origination and decrease over time. To achieve this reduction in skin-in-the-game and incentives, the optimal contract mandates smooth, time-decreasing payments to the agent. Therefore, the optimal contract can be implemented by requiring the lender to initially retain a share of the loan that it gradually sells to investors. Retention and selloff dynamics thus reflect the underlying moral hazard in screening and monitoring, and vice versa, in line with recent empirical findings. In particular, the underlying moral hazard problem shapes retention and selloff dynamics, consistent with the evidence in Chen, Lee, Neuhann, and Saidi (2023), Haque, Mayer, and Wang (2023), and Jiang, Kundu, and Xu (2023) that reduced moral hazard in loan syndication is associated with lower retention by the lead arranger and more loan sales.[3] And, conversely, monitoring increases with the loan share of the lender, as documented in Gustafson, Ivanov, and Meisenzahl (2021), and decreases as the lender sells its share.

Our model generates initial retention levels and loan sale dynamics that are consistent with those documented in the empirical literature. For example, in line with the evidence in Blickle et al. (2022), (*i*) most loan sales occur relatively shortly after origination, and (*ii*) the lender may sell the entire loan shortly after origination.[4] The latter scenario prevails when the benefits of monitoring (relative to its costs) are limited, that is, when the lender cannot add much value via the monitoring.

The model also allows us to examine the effects of loan and lender characteristics, such as loan maturity, borrower quality, or lender cost of capital, on retention dynamics. Higher intrinsic (pre-screening) credit risk implies earlier default and thus both a shorter time period

[3]Exploiting plausibly exogenous shocks to the severity of moral hazard in loan origination, Chen et al. (2023) and Jiang et al. (2023) show that, as the lender's (lead arranger's) moral hazard in screening and monitoring is alleviated, the lender retains a lower loan share and sells more of the loan to non-bank intermediaries. Haque et al. (2023) show for U.S. syndicated loans that the presence and actions of private equity (PE) sponsors reduce the necessity of bank monitoring for PE-backed loans, thus allowing the lead arranger retain a lower loan share and to sell more loan shares to non-bank intermediaries.

[4]For instance, our model can generate an initial retention level of 25%, in line with Sufi (2007), together with a full loan sale within about 100 days after origination, in line with Blickle et al. (2022).

over which the lender is exposed to the loan and a lower intrinsic loan value. As such, higher intrinsic credit risk makes it both more difficult and less interesting financially to incentivize screening and monitoring, leading to lower initial retention by the lender and faster loan sales after origination as part of the optimal contract. Thus, while Ivashina (2009)—respectively Wang and Xia (2014) and Gustafson et al. (2021)—document a negative relation between screening—respectively monitoring—and credit risk, our results point to a two-way causality. Not only do screening and monitoring reduce credit risk, but intrinsic credit risk also dampens monitoring and screening efforts. Through this mechanism, our model provides a rationale for the segmentation observed in credit markets, whereby lenders (such as banks) that exert high screening and monitoring typically finance high-quality borrowers.

We also show that a higher cost of capital for the lender implies greater gains from trade, so that the lender retains a lower share in the loan, sells it faster and is more likely to sell the entire loan, in line with the empirical findings of Irani and Meisenzahl (2017) and Irani et al. (2021). Lastly, shorter loan maturity reduces the amount of time that the lender is exposed to loan performance (but without reducing intrinsic loan value), which weakens its incentives to screen and increases credit risk. To counteract this effect, the optimal contract front-loads incentives by increasing initial retention. Therefore, the model predicts that short maturity debt should feature higher initial retention and monitoring incentives, but also higher selloff speed and lower screening, relative to long maturity debt.

An important question for empirical research is whether the share of the loan originator 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 but not for monitoring effort because subsequent loan sales undo monitoring incentives. Empirical measures for monitoring should take into account the selloff dynamics after origination. In particular, monitoring incentives should increase with the incentives of the lead bank, as captured by the contemporaneous lead share, in line with evidence

in Gustafson et al. (2021). We additionally show that while selloff speed is monotonic in the level of monitoring effort, it is non-monotonic in 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.

Next, we study various extensions of our baseline model. First, we consider that the lender originates a portfolio of two loans. Instead of retaining shares in each of the individual loans, the lender optimally creates tranches of the loan portfolio, akin to securitization. The loan portfolio is tranched into an equity (junior) tranche, which is wiped out after the first loan defaults, and a senior tranche, which only takes losses when the entire loan portfolio defaults. Optimal screening and monitoring incentives are provided by having the lender retain a share of the equity tranche that is gradually sold after origination, a pattern empirically observed for mortgage loans (Begley and Purnanandam, 2016).

Second, we consider repeated lender-investor interactions where the process by which the lender makes a loan and sells it to investors is repeated. The prospect of collecting payoffs from repeated loan origination provides screening and monitoring incentives, which allows the lender to retain a lower share of the loan at and after origination, in line with evidence in Gopalan, Nanda, and Yerramilli (2011). Thus, with repeated interactions, the lender may have strong incentives to screen and monitor so that loans need not perform poorly even when the lead lender sells its share relatively quickly after origination, thereby rationalizing the evidence in Blickle et al. (2022). Additionally, while our baseline analysis solves for the optimal retention dynamics under full commitment, we show that repeated lender-investor interactions can generate such commitment. Intuitively, originating the loan and selling it to outside investors is profitable for the lender. If these gains are sufficiently large and deviating from the retention path stipulated in the contract implementation hampers future loan sales, the lender will have sufficient incentives to comply with the prescribed retention path.

Third, in some applications of credit securitization (e.g. 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. corporate loans), they are undertaken by the same entity. To understand the consequences of separation, we consider a model variant in which two otherwise identical agents, respectively, screen and monitor loans and, to have adequate incentives, retain a stake in the 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 screening and monitoring incentives. On the contrary, 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 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, relative to markets where these two tasks are bundled and undertaken by the same entity, as is common for syndicated loans.

Our paper relates to the extensive 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 distinguish between monitoring after loan origination and screening at origination and cannot investigate the dynamics of incentives and loan sales and their effects on credit risk and loan value. Following early contributions by Sufi (2007) and Ivashina (2009), a growing empirical literature examines the effects of the share 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)). Most of these studies proxy skin in the game by initial retention. This literature has recently focused on loan sales after origination and their effects on incentives and credit risk (Lee, Liu, and Stebunovs, 2022; Blickle et al., 2022; Chen et al., 2023).

Our paper contributes to this literature mainly in two ways. First, we highlight the key role of the originator's contemporaneous loan share for screening and monitoring incentives, and rationalize loan sales after origination as part of an optimal contract between loan originators and outside investors. Second, we shed light on the complex relationship between screening and monitoring and the originator's skin in the game. In particular, we demonstrate

that both initial retention and selloff speed determine incentives and that incentives are best captured by the share of the agent when they exert effort both for screening—initial originator share—and for monitoring—contemporaneous originator share.

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 (2021a) characterize optimal monitoring in dynamic settings but do not focus on optimal contracts. In a related paper, Hartman-Glaser, Piskorski, and Tchistyi (2012) study optimal securitization and screening of mortgages under moral hazard. In their model, the optimal contract features a single payout to the agent when sufficient time has elapsed after the 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 in several ways. First, while in practice corporate loans are both screened and monitored, our paper is the first to model screening and monitoring in a unifying framework. We show that the combination of screening and monitoring moral hazard implies that the optimal contract between the lender and investors 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). Notably, unlike these theories, our model can generate retention and selloff dynamics that mirror the patterns documented in recent empirical studies. Second, the model allows us to examine the effects of loan and lender characteristics on retention dynamics. This allows us to rationalize recent empirical findings and to generate new testable predictions regarding optimal retention and the dynamics of loan sales by originators and their effects on loan performance.

1 Model setup

Time t is continuous and defined over $[0, \infty)$. A lender (the agent) originates a loan that can be sold to competitive outside investors (the principal). In the model's key application, namely syndicated lending, the lender represents the lead arranger, while investors represent other banks in the syndicate or institutional investors (e.g., CLOs or loan market mutual funds) who buy loans in the secondary market. In the baseline model, the loan has infinite maturity. An equivalent interpretation is that the loan has finite maturity, but is rolled over every time it matures until default. Section 4.3 shows that the implications of the model are not affected if loans have finite maturity and are not rolled over.

1.1 Screening, monitoring, and default risk

The loan promises a constant flow payoff (coupon payment) normalized to 1 up to its default, which occurs at random time τ. The liquidation value of the loan at default is normalized to zero for simplicity, as, e.g., DeMarzo and He (2021). 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 $t \geq 0$, in that

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

In this equation, $\Lambda > 0$ captures the intrinsic quality (default intensity) of the loan. Screening effort q captures the lender's due diligence and screening of the borrower prior to loan origination, where a higher q corresponds, e.g., to more information collected and processed during the due diligence process and, thus, to lower levels of default risk.[5]

Monitoring effort $(a_t)_{t\geq 0}$ captures the lender's post-origination due diligence and moni-

[5]To make our baseline analysis tractable, we model the impact of screening effort on default risk λ_t in reduced form as in Hartman-Glaser et al. (2012), Malamud et al. (2013), and Hoffmann et al. (2021). Appendix B.6 provides a micro-foundation of the loan origination process and the impact of screening effort on default risk in which screening effort allows the lender to distinguish good from bad borrowers, thereby reducing the loan's default risk. The model solution and analysis are similar, but less tractable.

toring, which can take various forms (see, e.g., Gustafson et al. (2021) and Heitz, Martin, and Ufier (2022) for direct evidence on bank monitoring). For instance, monitoring could capture a lender's on-site inspections of the borrower, third-party appraisals (a third party is hired by the lender to conduct an audit/inspection of the borrower), the active request and verification of the borrower's financial or collateral information, or the monitoring and enforcement of loan covenants. The lender's monitoring effort may in practice curb borrower moral hazard, prevent borrower risk-taking, and more generally improve the likelihood that the lender is repaid. In the following, we assume that monitoring effort a_t reduces the default intensity λ_t. This modeling assumption is in line with empirical evidence in Heitz et al. (2022) that active monitoring by the lender (e.g., via on-site inspections) reduces default risk and in Blickle, Parlatore, and Saunders (2023) that both pre-origination screening and post-origination monitoring improve loan performance (i.e., reduce default risk).

Screening and monitoring efforts are bounded in that $q \in [0, \bar{q}]$ and $a_t \in [0, \bar{a}]$ with $\Lambda > \bar{a} + \bar{q}$. The bounds $\bar{a}$ and $\bar{q}$ are necessary to ensure that the instantaneous default probability λ_t is well defined and positive. Unless otherwise mentioned, we focus on parameter configurations that lead to optimal efforts $a_t \in [0, \bar{a})$ and $q \in [0, \bar{q})$, so that the upper bounds do not bind and the model solution, as well as contract dynamics, do not depend on the exact values of $\bar{a}$ and $\bar{q}$. We discuss formally binding upper bounds in Appendix B.7.

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 and are not contractible, giving rise to moral hazard. We do not impose any restrictions on the relation between screening and monitoring. In particular, we do not make any assumptions about 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.[6] If the lender decides to shirk on either task, the loan will have a higher default rate. Although both reduce the risk of default, it is important to note that screening occurs only once, when the loan is originated at time $t = 0$, whereas monitoring occurs at any point in time $t \geq 0$ up

[6] We can allow screening and monitoring to be complements or substitutes in reducing default risk with little effect on the model solution and analysis by assuming for example that $\lambda_t = \Lambda - q - a_t - \alpha q a_t$. See Appendix B.1 for a theoretical analysis of this model variant, Appendix B.6 for its micro-foundation, and Section 2.2.4 for a numerical analysis.

to default. Furthermore, the effect of screening is more persistent than that of monitoring, where we consider for tractability that the impact of monitoring is purely transitory.

1.2 Gains from trade and loan sales

Both the principal and the agent are risk neutral.[7] 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 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 long-term 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 sets out a profit-sharing rule for the 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 specifies the monitoring effort $\hat{a}_t$ (for all $t \geq 0$) and the screening effort $\hat{q}$. We focus on incentive compatible contracts that induce actual monitoring and screening efforts to coincide with contracted monitoring 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 under the contract $\mathcal{C}$ must at any time exceed their outside option, which we normalize to zero. The principal and the agent are able to fully commit to the transfer rule $(dC_t)_{t\geq 0}$ stipulated by the optimal contract as long as it meets their limited liability constraint.[8] We do not impose explicit constraints on the transfers dC_t after time zero, but show later that optimal transfers satisfy $dC_t \geq 0$ for $t > 0$.

[7] Alternatively, 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.

[8] Section 4.4 shows how commitment can arise through repeated originator-investor relations.

1.3 Contracting problem

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

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

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. Once the loan defaults at time τ, there are no more coupon payments and the game ends, so both the principal's and the agent's continuation payoff fall to zero.[9] Thus, $dC_t = 0$ 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 competitive 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

$$P_t = \mathbb{E}_t\left[\int_t^\tau e^{-r(s-t)}(1 - c_s)ds\right] = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}(1 - c_s)ds. \tag{3}$$

In equation (3), 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 $dC_{0^-} = P_0$. Under the contract $\mathcal{C}$, the agent's continuation payoff W_t at time $t \geq 0$ is given by the present value of the future payments adjusted for the cost of effort:

$$W_t := \mathbb{E}\left[\int_t^\tau e^{-\gamma(s-t)}\left(c_s - \frac{\phi a_s^2}{2}\right)ds\right] = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du}\left(c_s - \frac{\phi a_s^2}{2}\right)ds. \tag{4}$$

W_t captures the value of the lender's stake in the loan. The limited liability constraints of

[9] After default at time τ, the loan is worth zero and so is the sum of agent's and principal's payoff. Due to limited liability, neither the agent nor the principal can have negative payoffs and, because their payoffs add up to zero, it follows that $dC_t = 0$ for $t > \tau$.

the lender and investors are then formally defined as $W_t \geq 0$ and $P_t \geq 0$ for any $t \geq 0$.[10]

As investors are competitive, the lender can extract all the surplus and therefore chooses the security that maximizes total surplus $F_{0^-} := W_{0^-} + P_0$ at time $t = 0^-$. That is, the lender solves

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

taking into account its own moral hazard problem (i.e., incentive compatibility constraints) and the limited liability constraints $W_t, P_t \geq 0$ for any $t \geq 0$.

Because P_t in equation (3) and W_t in equation (4) 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 $(a_t)_{t\geq 0}$. To begin with, let us fix screening effort at q and analyze monitoring incentives given q. Due to limited liability, the agent only loses its claim to future payments, i.e., its continuation payoff W_t, at the time of default. With its 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

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

[10] That is, if $W_t < 0$ or $P_t < 0$, the lender or investor would be better off leaving the contractual relationship and enjoying their outside option (normalized to zero). At time $t = 0^-$, the limited liability constraint for the lender implies $W_{0^-} = W_0 - \frac{\kappa q^2}{2} \geq 0$, that is, the expression for the agent payoff in (2) is positive.

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

$$a_t = \frac{W_t}{\phi}. \tag{6}$$

The incentive constraint for monitoring effort (6) shows that incentive compatibility requires $\hat{a}_t = a_t = \frac{W_t}{\phi}$ for all $t \geq 0$. Granting the lender a higher stake W_t increases its exposure to default risk and monitoring incentives, but is costly due to its 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 the screening effort q chosen at time $t = 0^-$. The agent chooses q to maximize W_{0^-} which is the value of its claim after screening is chosen, $W_0(q)$, net of the screening effort cost, $\frac{\kappa q^2}{2}$:

$$\max_{q\in[0,\bar{q}]} \left(W_0(q) - \frac{\kappa q^2}{2} \right). \tag{7}$$

Denote by V_t the agent's gain from a marginal increase in q measured from time t onward:

$$V_t = \frac{\partial}{\partial q} W_t(q). \tag{8}$$

We can use V_0 to write the first-order condition solving (7) for the optimal screening effort:

$$q = \frac{V_0}{\kappa}. \tag{9}$$

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, the equation (9) describes incentive compatibility for the screening effort, in that $q = \hat{q} = \frac{V_0}{\kappa}$.

While V_0 determines screening effort, the optimal contract will depend on the whole path of V_t beyond $t = 0$. Notably, we show later that V_t becomes a state variable for the dynamic optimization problem of the lender because the optimal long-term contract takes into account how time-t incentives affect screening incentives at time $t = 0$. To characterize V_t and V_0,

we differentiate the integral representation of W_t in (4) under optimal a_t to obtain:[11]

$$V_t = \int_t^\infty (s-t) e^{-\gamma(s-t)-\int_t^s \lambda_u du} \left(c_s - \frac{\phi a_s^2}{2} \right) ds = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du} W_s ds. \tag{10}$$

Note that both screening and monitoring incentives are provided by exposing the agent to loan performance via $W_t > 0$. Higher W_t exposes the agent more strongly to loan performance and therefore motivates screening. Furthermore, a higher W_t increases monitoring a_t, which delays default and strengthens screening incentives measured by V_t. Equation (10) 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., $V_t = D_t W_t$. This decomposition captures the intuition that screening incentives are the strongest if the exposure W_t to the loan is large and has a high duration D_t. This creates a trade-off as late payments increase duration but decrease value. The determination of screening incentives must therefore resolve the tension between duration and value.

Next, we characterize the dynamics of the agent's monitoring and screening incentives W_t and V_t. We can differentiate (4) with respect to time and obtain

$$\dot{W}_t := \frac{dW_t}{dt} = (\gamma + \lambda_t) W_t + \frac{\phi a_t^2}{2} - c_t. \tag{11}$$

Similarly, differentiating V_t in (10) with respect to time t, we obtain the dynamics of V_t:

$$\dot{V}_t := \frac{dV_t}{dt} = (\gamma + \lambda_t) V_t - W_t. \tag{12}$$

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

[11] When differentiating W_t, we can ignore the effect on a_t due to the envelope theorem. Also, because screening effort q is neither observable nor contractible, an unobserved change in q cannot affect the contracted flow payments c_t. A derivation of (10) is provided in the proof of Proposition 2.

Lemma 1. *Suppose that the model parameters satisfy*

$$\kappa > \frac{2}{(r+\Lambda-\bar{a}-\bar{q})(\gamma+\Lambda-\bar{a}-\bar{q})^2} + \frac{1}{\phi(r+\Lambda-\bar{a}-\bar{q})^2(\gamma+\Lambda-\bar{a}-\bar{q})^3}. \tag{13}$$

Incentive conditions (6) *and* (9) *hold and uniquely pin down monitoring and screening efforts. Incentive conditions* (6) *and* (9) *are sufficient and the first-order approach is valid.*

Throughout the paper, we assume that condition (13) is met and that

$$\kappa > \frac{\phi\bar{a}}{\bar{q}(\gamma+\Lambda-\bar{a}-\bar{q})}, \tag{14}$$

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 screening q. Note that with observable q, unobservable actions a_t have immediate rather than persistent effects. Additionally, absent default, our environment remains constant over time. We thus 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 $\dot{W}_t = \dot{a}_t = 0$, $W_t = W = W^B(q)$, and $a_t = a^B(q)$ for all t. Inserting $\dot{W}_t = 0$ into (11) yields

$$c = (\gamma+\Lambda-a-q)W + \frac{\phi a^2}{2}. \tag{15}$$

Given screening q and monitoring a, the default rate is constant and equal to $\Lambda - a - q$, and the price of the security paying flow payouts $1-c$ to investors becomes

$$P^B(q) = \frac{1-c}{r+\Lambda-a-q}. \tag{16}$$

Equation (15) 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. Next, note that given q, optimal monitoring effort a (and equivalently optimal deferred compensation $W = \phi a$) is chosen to maximize total surplus after screening, $F^B(q) = P^B(q) + W$. Using equations (15) and (16), we thus get that the lender solves

$$F^B(q) = \max_{W \in [0, F^B(q)]} \left(\underbrace{\frac{1}{r + \Lambda - a - q}}_{\text{Market value}} - \underbrace{\frac{(\gamma - r)W}{r + \Lambda - a - q}}_{\text{Agency cost}} - \underbrace{\frac{\frac{\phi a^2}{2}}{r + \Lambda - a - q}}_{\text{Monitoring cost}} \right), \tag{17}$$

subject to $a = W/\phi$ (incentive compatibility) and $W \in [0, F^B(q)]$ (limited liability). Equation (17) shows that the surplus $F^B(q)$ consists of the present value of the loan payments 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 (17) yields optimal levels of monitoring effort

$$a^B(q) = \max \left\{ \frac{F^B(q) - (\gamma - r)\phi}{\phi}, 0 \right\}, \tag{18}$$

and $W^B(q) = \phi a^B(q) < F^B(q)$, given a level of screening q. Using (10), we can also calculate

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

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 optimize $F^B(q)$ over q to determine optimal screening in this second-best benchmark: $q^B = \arg\max_{q \in [0, \bar{q}]} \left(F^B(q) - \frac{\kappa q^2}{2} \right)$. We summarize our findings in the following proposition.

Proposition 1 (No moral hazard over screening)**.** *Suppose that screening effort q is contractible so that there is no moral hazard with respect to screening. At the optimum, monitoring effort $a^B(q)$, payouts $c^B(q)$, and deferred payouts $W^B(q)$ $(< F^B(q))$ are constant over time and jointly characterized by (6), (15), and (17) for any choice of q. Optimal monitoring effort $a^B(q)$ increases with q. Optimal screening effort 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. We first fix screening q and solve the continuation problem for $t \geq 0$. We then determine optimal screening $q = q^*$, taking into account the solution to the continuation problem.

Given monitoring a and screening q, we can write the total surplus at time t as[12]

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

with $W_t = \phi a_t$. The time-0^- optimization can then be written via the Lagrangian:

$$\mathcal{L}_{0^-} = F_{0^-} + \ell(\kappa q - V_0),$$

where ℓ is the Lagrange multiplier for the screening incentive constraint $\kappa q = V_0$. Maximizing the Lagrangian for each time t while taking into account the monitoring incentive constraint (6) yields that optimal effort a_t, if interior, satisfies the first order condition:

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

Therefore, we have that when a_t is interior

$$a_t = \frac{\overbrace{F_t}^{\text{Reduction of default risk}} - \overbrace{(\gamma - r)\phi}^{\text{Agency costs}} \overbrace{-\ell e^{-(\gamma - r)t}(V_t + \phi)}^{\text{Screening incentives}}}{\underbrace{\phi}_{\text{Direct cost}}} \quad \wedge \quad \frac{F_t}{\phi}, \tag{21}$$

[12] For a derivation, take $F_t = P_t + W_t$ in the first line of (20) and take the derivative with respect to t:

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

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

where $\min\{x, y\} = x \wedge y$ and where we account for the possibility that the principal's limited liability constraint binds (in which case $W_t = \phi a_t = F_t$). See Appendix A.3.3 for a derivation of this result. The intuition for (21) is that monitoring reduces the probability of default but comes at additional direct and agency costs. In addition, in a long-term contract, the optimal choice of effort at time $t > 0$ takes into account its effect on screening incentives at origination, as captured by $-\ell e^{-(\gamma-r)t}(V_t + \phi)$, which distorts optimal monitoring away from the benchmark level with contractible screening in (18). As the agent is relatively more impatient and $\gamma > r$, this effect, however, vanishes over time. Thus, optimal monitoring a_t and, consequently, V_t, W_t, and F_t approach the respective levels of the benchmark with observable screening as time t tends to ∞, in that

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

For times $t < \infty$, V_t affects the optimal choice of monitoring effort in (21), and thus becomes a relevant state variable in the dynamic optimization of total surplus.

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 state 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:

$$\begin{aligned} rF(V,W) = \max_{a,c}\Bigg\{ & 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V,W) \\ & + F_V(V,W)((\gamma+\lambda)V - W) + F_W(V,W)\left((\gamma+\lambda)W + \frac{\phi a^2}{2} - c\right)\Bigg\}, \end{aligned} \tag{22}$$

where $F_V(V,W) = \frac{\partial F(V,W)}{\partial V}$ and $F_W(V,W) = \frac{\partial F(V,W)}{\partial W}$, and where we have used the dynamics of W and V given in (11) and (12).[13] Equation (22) is solved subject to the incentive condition (6), the limited liability constraints, and the conjecture that payouts to the lender

[13] For a derivation, conjecture that $F_t = F(V_t, W_t)$, so $\dot{F}_t = F_V(V_t, W_t)\dot{V}_t + F_W(V_t, W_t)\dot{W}_t$. Differentiate (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 (22) after inserting $\dot{F}_t = F_V(V_t, W_t)\dot{V}_t + F_W(V_t, W_t)\dot{W}_t$ and $F_t = F(V_t, W_t)$.

are smooth, in that $dC = cdt$. Note that it is always possible to stipulate that the lender receives an incremental payout of Δ dollars,[14] which leaves V unchanged but changes W by $-\Delta$ dollars. 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 (22) as a control variable. Optimal payouts to the lender are then defined as the residual that implements the optimal W, as we show in Section 3.1.

As we do not impose any constraints on the payout rate c and it is always possible to increase or decrease c, the optimality of payouts c requires the first order condition

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

to hold. Substituting $F_W(V,W) = 0$ back into (22) yields

$$rF(V) = \max_{a\in[0,\bar{a}],W} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)\big((\gamma+\lambda)V - W\big) \right\}, \tag{23}$$

where (with a slight abuse of notation) $F(\cdot)$ is a function of V only and W is a control. Equation (23) is solved subject to the incentive condition for monitoring effort (6), i.e., $W = \phi a$, and the principal's and the agent's limited liability conditions, i.e., $W \in [0, F(V)]$.

As $t \to \infty$, the state variable V_t approaches $V^B(q)$ which is defined in (19). Expressed in terms of the state variable V, equation (23) is solved subject to the boundary condition

$$\lim_{V \to V^B(q)} F(V) = F^B(q). \tag{24}$$

We assume that a unique, continuously differentiable solution $F(V)$ to (23) subject to (24) exists. 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 $\lim_{t\to\infty} \dot{V}_t = 0$. Thus, the state space can be characterized by the interval $(V^B(q), V_0]$. The value function is downward sloping, with $F'(V) < 0$ for $V \in (V^B(q), V_0]$. We also show that the value function is strictly concave.

Having characterized the model solution for $t \geq 0$ and given q, we are now in a position

[14] If payouts to the lender are not smooth, then it follows similar to (11) 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, $dW = -\Delta$.

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 (9):

$$q^* = \arg\max_{q\in[0,\bar{q}]} \left(F(V_0) - \frac{\kappa q^2}{2} \right) \quad \text{s.t.} \quad V_0 = \kappa q. \tag{25}$$

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 (4) and (10) respectively, and have dynamics given by (11) and (12) 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 (23) subject to boundary condition (24).*

2. *Optimal monitoring is characterized by the maximization in (23) subject to (6). Optimal screening effort $q = q^*$ is characterized in (25).*

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

Finally, note that the optimal contract is designed to maximize the total surplus for the lender and investors, given that the loan is originated. As such, the optimal contract would not change if we modelled the initial decision to extend a loan of size K to the borrower. In that case, the optimal contract would be designed to maximize $F_{0^-} - K = F_0 - \frac{\kappa q^2}{2} - K$. While the exact value of K would affect the initial surplus, subject to the participation constraint $F_{0^-} - K \geq 0$, it would not affect the contract dynamics.[15]

2.2.3 Contract Dynamics

Figure 1 provides a numerical example of the optimal contract. For the numerical analysis,

[15] Section B.6.8 provides a micro-foundation for loan size.

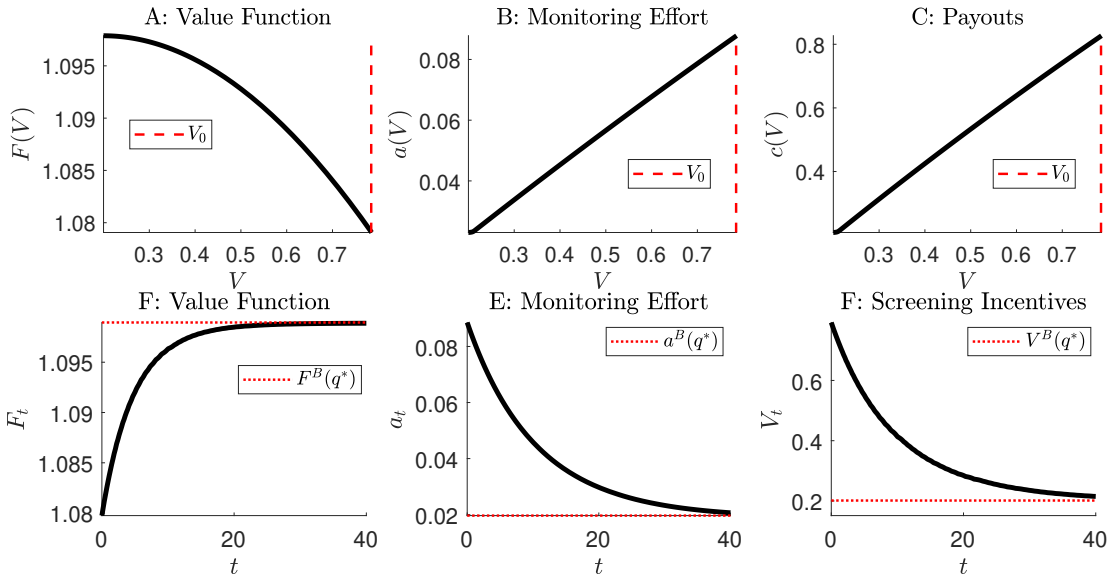

Figure 1: **Total surplus $F(V)$, monitoring $a(V)$, and the agent's flow payouts $c(V)$.** In the upper panels, the vertical dashed red line denotes the V_0. In the lower panels, the horizontal dotted red line denotes the benchmark levels that are attained in the limit $t \to \infty$.

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 $1/(\Lambda + r) = 1$.[16] In addition, we set $\gamma = 0.1$ and $\phi = \kappa = 9$ to generate the desired trade-offs. Lastly, we pick $\bar{a} = 0.125$ and $\bar{q} = 0.24$ to satisfy conditions (13) and (14). Our parameter choices imply 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.

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 and increase with V, i.e., decrease over time since $\dot{V} < 0$. 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 $W_t = \phi a_t$, we do not plot it separately.) Observe that V_t, W_t, and a_t decrease over time

[16] 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).

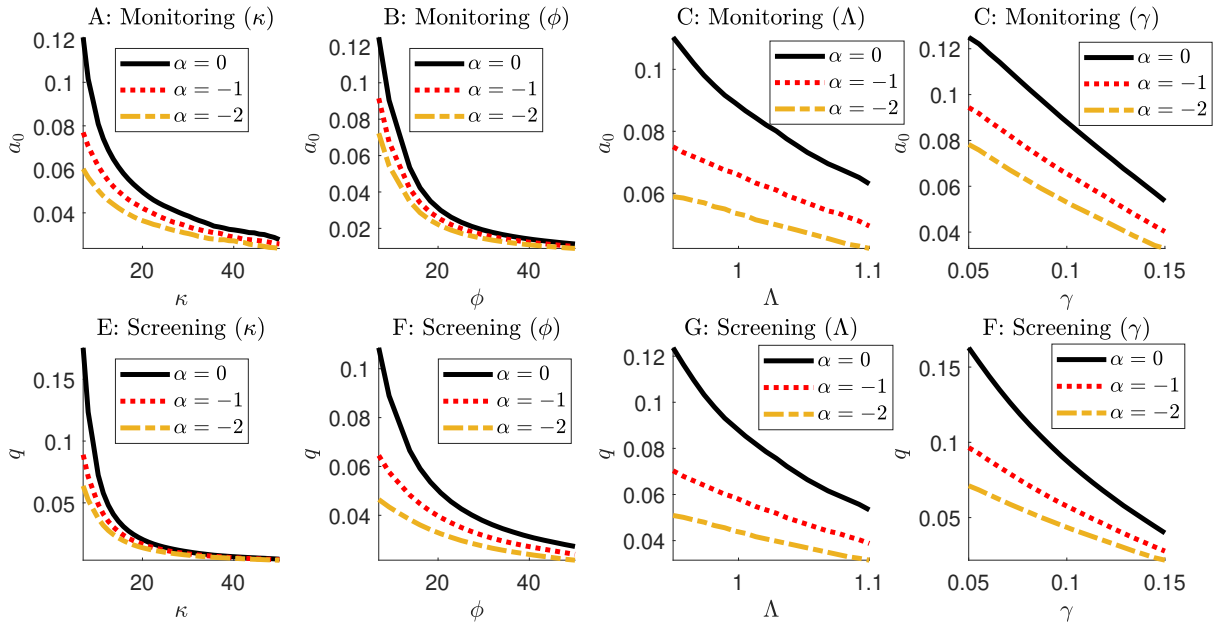

Figure 2: **Comparative Statics.** This figure plots monitoring effort a_0 for $\alpha = 0$ (solid black line) and screening effort q^* against the parameters ϕ, κ, and Λ, for $\alpha = -1$ (dotted red line), and $\alpha = -2$ (dashed yellow line). We use our baseline parameters.

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 $F^B(q^*)$ in the long run.

2.2.4 Determinants of Incentives

We now study the determinants of incentives by performing a comparative static analysis of monitoring and screening efforts with respect to exogenous model parameters. The key finding of this section is that due to moral hazard, screening and monitoring endogenously arise as complements. To underscore the robustness of this result, we consider a generalization of

our baseline model in which the loan default intensity is given by:

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

When $\alpha > 0$ ($\alpha < 0$), screening and monitoring are complements (substitutes) in reducing default risk. In the baseline model, we have $\alpha = 0$. That is, we do not make any assumptions on whether screening and monitoring efforts are substitutes or complements. The solution for this model variant is analogous to that of the baseline model as shown in Appendix B.1. A micro-foundation of this default intensity can be found in Appendix B.6.

Figure 2 plots initial monitoring a_0 (which proxies for overall monitoring) and screening q as functions of the cost of screening κ, the cost of monitoring ϕ, intrinsic credit risk Λ, and lender cost of capital γ for $\alpha = 0$, $\alpha = -1$, and $\alpha = -2$. Panels A, B, E, and F of Figure 2 show that monitoring effort a_t and screening effort q decrease with both the 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. 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 monitoring (incentives), leading to lower screening (incentives). Notably, screening and monitoring endogenously arise as complements for incentive purposes even for negative values of α, that is, when assuming that screening and monitoring are substitutes in reducing credit risk and absent moral hazard.[17]

Panels C and G of Figure 2 illustrate that a decrease in the intrinsic quality of the loan, as reflected by the higher baseline default intensity Λ, leads to a decrease in monitoring and screening. That is, our paper suggests a two-way relation between credit risk and lenders' screening and monitoring. Notably, a lower credit quality leads to laxer monitoring and

[17] The complementarity of screening and monitoring may vanish for sufficiently large negative values of α. Obviously, the complementarity is stronger for positive values of α.

screening, which in turn exacerbates credit risk. Indeed, a higher rate of default Λ implies a lower expected duration for the loan and the agent's payments, which in turn makes it more costly to provide screening incentives. Thus, for larger values of Λ, it becomes optimal to reduce screening incentives which also leads to a reduction of monitoring incentives.

Finally, Panels D and H of Figure 2 show that screening and monitoring efforts decrease with γ, as it becomes more costly to delay payouts to the lender and to provide incentives.[18]

3 Dynamic Retention and Loan Sales

3.1 Contract Implementation via Dynamic Retention

This section shows that the optimal contract can be implemented by having the lender keep 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 (for $t \geq 0$), the lender (progressively) sells off its stake so that β_t 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

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

at any time $t \geq 0$ where credit risk is captured 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 coupon payments from the loan. In addition, selling the loan at rate $-d\beta_t$ yields trading revenues $-d\beta_t L_t$. Therefore, matching the payoffs to the payouts $c_t dt$ of the optimal contract requires that

$$\beta_t dt - d\beta_t L_t = c_t dt. \tag{27}$$

We can solve (11) to get

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

[18] This is consistent with the evidence in Purnanandam (2011) that securitization reduces screening and performance in mortgage markets and that this effect is more pronounced for more capital-constrained banks.

As payouts to the lender are smooth and positive for $t > 0$, retention will be smooth too, so $d\beta_t = \dot{\beta}_t dt$. Equations (28) and (27) then imply the ODE:

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

This equation is solved subject to $\lim_{t\to\infty} \beta_t = c^B = c^B(q^*)$, where $c^B(q)$ is the constant payout level in the limit $t \to \infty$ (or, equivalently, $V_t \to V^B(q)$) characterized in Proposition 1 under optimal screening $q = q^*$ (see also Appendix A.4).

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).*

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. 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 stationary with constant monitoring $a^B(q) = W^B(q)/\phi$ and constant payouts $c^B(q)$ up to default. The contract can then be implemented by having the agent retain a constant share of the loans $\beta^B(q) = c^B(q)$.

In the limit $\phi \to \infty$,[19] monitoring is prohibitively costly, so both contracted and actual monitoring equal zero and mechanically there is no moral hazard over monitoring.[20] Thus, the default intensity equals $\Lambda - q$ and is constant over time. Without moral hazard over monitoring, the optimal contract stipulates constant payouts $c_t = 1$ up to time τ^0 (finite and endogenous). At time τ^0, the agent receives, in addition, a lumpy payout $dC_{\tau^0} > 0$. This contract maximizes the agent's exposure to loan performance before time τ_0, while respecting the principal's limited liability. The implementation of the optimal contract then requires

[19] Likewise, one could consider the case $\phi = 0$ so that $a_t = \bar{a}$ without moral hazard. This leads to default intensity $\Lambda - \bar{a} - q$. The model with $\phi = 0$ is isomorphic to the limit $\phi \to \infty$ upon replacing Λ with $\Lambda - \bar{a}$.

[20] An increase in ϕ relaxes the parameter condition (13) but tightens (14) (which in fact cannot hold in the limit). We therefore can stipulate $\bar{a} = \bar{\chi}/\phi$ for appropriate constant $\bar{\chi} > 0$ (large enough to ensure effort is interior) so that (14) is met in the limit $\phi \to \infty$. This is merely a technical assumption and does not affect any of the conclusions, as a_t tends to zero regardless for $\phi \to \infty$.

the lender to retain the entire loan until time τ^0, at which point it fully sells the loan to investors. As an alternative to the limit argument, Appendix B.8 solves the model when there is no moral hazard over monitoring (a_t is observable and contractible) but $\phi < \infty$, and shows that the outcomes in this model variant are similar, i.e., the agent retains the entire loan at origination and does not sell up to some time τ^0. We thus have that:

Proposition 4. *The following holds:*

1. *When there is no moral hazard over screening, the optimal contract stipulates after time $t = 0$ constant payouts up to default at rate c^B. The optimal contract can be implemented by having the agent retain a constant fraction of the loan $\beta^B = c^B$.*

2. *When there is no moral hazard over screening, there exists a finite time $\tau^0 \in (0, \infty)$ such that the optimal contract stipulates smooth payouts at rate $c_t = 1$ for all times $t \in (0, \tau^0)$. At time τ^0, the optimal contract stipulates a (strictly positive) lumpy payout $dC_{\tau^0} > 0$ to the agent. The optimal contract can be implemented by requiring the lender to retain the entire loan until time τ^0, at which point it fully sells the loan to investors.*

3.2 Application to Syndicated Loans

While our model applies to credit markets broadly, we focus in what follows on the market for syndicated loans in which loan sales are common, as shown for example by Drucker and Puri (2009); Irani et al. (2021).[21] We start this section with a brief discussion of some of the institutional details of the syndication process and how they relate to our model.

The syndication process, which is described in greater detail in Bruche, Malherbe, and Meisenzahl (2020), consists broadly of three stages. In the first stage ("origination stage"), the lead lender—also referred to as the lead bank or lead arranger—matches with a borrower and conducts due diligence (screening). Provided the outcome of the screening process is positive, the lead lender and co-investors (other banks in the syndicate) jointly commit the

[21] Loan sales and securitization are common in other markets too and affect lender incentives. For instance, Purnanandam (2011) and Keys, Mukherjee, Seru, and Vig (2010) find that securitization and the originate-to-distribute model have led to reduced screening in the market for mortgages prior to the subprime crisis.

loan to the borrower, with loan terms being determined based on the screening outcome.[22]

In the second stage ("book running" or "primary market"), the deal is marketed to outside investors, which can be other banks or institutional investors (e.g., CLOs or loan market mutual funds). During this stage—which lasts on average 46 days (Bruche et al., 2020)—outside investors may buy loan shares right away or commit to buying loan shares in the secondary market.[23] That is, during the primary market stage, the lead arranger gradually reduces its exposure to the loan by engaging in loan sales or, alternatively, pre-committed loan sales (akin to a forward sale of the loan). During this primary market stage, the lead arranger is exposed to pipeline risk, i.e., the risk that it cannot sell the loan if investor demand dwindles, e.g., due to bad news about the borrower (not necessarily limited to actual default).[24] Broadly interpreted, the Poisson process dN_t captures such bad news. In the third stage ("secondary market"), the secondary market opens. Outside investors can then buy the loan and pre-committed sales can be executed.

In our model, the first stage runs from time $t = 0^-$ to time $t = 0$ and β_0 can be seen as the lead arranger's initial share of total credit commitment. Then, times $t > 0$ represent the second and third stages (i.e., primary and secondary market), during which the lender gradually reduces its exposure to the loan. Crucially, the implementation of the optimal contract via the lender's time-varying retention β_t allows us to map our model to the data. In particular, the empirical analog for β_t is the lead arranger's share which is reported at origination in the Dealscan database and over time in the Shared National Credit Registry.

Figure 3 plots the lender's share β_t against time t, both under our baseline parameters (Panel A) and when ϕ and γ are larger (Panel B). As time passes, the agent sells its stake β_t. Thus, our model generates optimal loan sales by the (lead) lender as part of the optimal lender-investor contract. Notably, as we argue next, the retention and loan sales dynamics qualitatively resemble the patterns observed in the data.

First, observe that the selloff speed, as captured by $-\dot{\beta}_t$ in Figure 3, decreases with

[22] While loan terms can be changed during the syndication process (e.g., due to lack of investor demand), it is very uncommon that the lenders renege the loan commitment.

[23] Typically and as discussed in Blickle et al. (2022), CLOs pre-commit to buy loan shares in the secondary market for tax reasons, instead of directly participating in the syndicate.

[24] Part of the pipeline risk is also borne by the borrower, as loan terms may be adjusted in response to weak demand from investors. Bad news may also annul pre-committed loan sales.

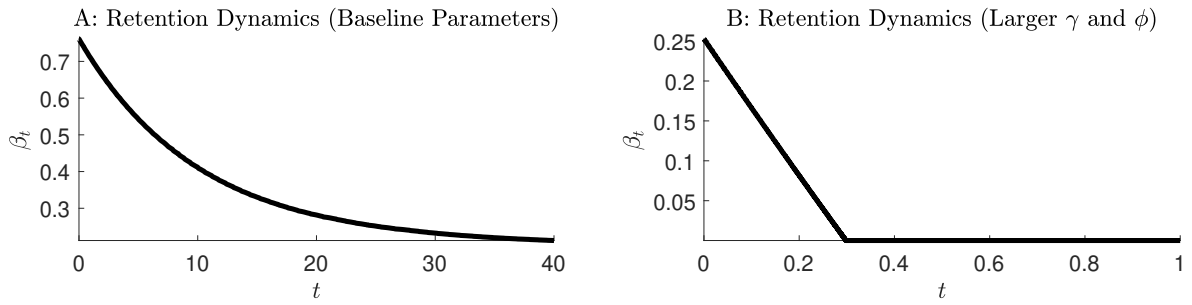

Figure 3: **Implementation of the optimal contract and per unit value of the loan.** Panel A uses our baseline parameters and Panel B sets $\gamma = 0.12$ and $\phi = 65$.

time t since origination. That is, in Panel A, β_t is convex and decreasing in t approaching β_B in the limit (the selloff speed tends to zero as $t \to \infty$), while in Panel B the selloff speed becomes zero at some point. The interpretation is that most of the loan sales occur (relatively) shortly after origination, consistent with the findings in Blickle et al. (2022) or Lee et al. (2022). In fact, under certain parameter conditions, the lender sells off its entire stake in finite time. In particular, our model is able to capture the selloff dynamics reported in Blickle et al. (2022) that in some cases (especially for Term B loans) the lender sells its entire stake relatively shortly after origination (e.g., within 100 days).[25] Panel B of Figure 3 plots the retained share when the cost of monitoring ϕ and lender's discount rate are larger than in the baseline. The lender retains initially $\beta_0 \approx 25\%$ of the loan—in line with the initial retention level reported in Sufi (2007)—and sells its entire stake after 104 days.

More generally, Corollary 1 shows analytically that when the lender's cost of capital γ or the cost of monitoring ϕ are sufficiently large, the lender sells off its entire stake in finite time. Thus, our model provides an explanation for some of the recent puzzling findings in the empirical literature on performance and skin in the game. Notably, our results suggest that lenders will only keep on their books loans that have a low holding cost (low γ) and to which they can add value through monitoring (low ϕ):

Corollary 1. *Under the implementation from Proposition 3, we have that*

[25] Blickle et al. (2022) report the number of days that it takes the lead arranger to sell its share from the point that the loan trades on the secondary market. However, as argued above, the lead lender effectively reduces its exposure to loan performance already in the primary market, which lasts for about 46 days on average. Thus, the quantities reported in Blickle et al. (2022) (e.g., days to selloff) represent a lower bound on the actual time that the lender is exposed to loan performance since committing the loan to the borrower.

1. *When the cost of monitoring is sufficiently high in that*

$$\phi > \max\left\{\frac{1}{(r+\Lambda-\bar{q})(\gamma-r)}, \frac{1}{(r+\Lambda-\bar{q}-\bar{a})(\gamma+\Lambda)}\right\}, \tag{30}$$

the lender sells off its entire stake in finite time. In this case, $\beta_0 > 0$ *and there exists time* $T \in (0,\infty)$ *such that* $\beta_t = 0$ *and* $W_t = 0$ *for* $t \geq T$*.*

2. *When the cost of monitoring is sufficiently high in that* $\phi < \frac{1}{(r+\Lambda)(\gamma-r)}$*, the lender never sells its entire stake, i.e.,* $\beta_t > 0$ *and* $W_t > 0$ *for all* $t \geq 0$*. Thus,* $\phi \geq \frac{1}{(r+\Lambda)(\gamma-r)}$ *is a necessary condition for selloff in finite time.*

Importantly, the sufficient condition (30) and the necessary condition ($\phi \geq \frac{1}{(r+\Lambda)(\gamma-r)}$) for full selloff do not depend on the cost of screening κ. Thus, holding Λ, r, and γ fixed, it is not possible to rule out that the loans which are sold off in finite time may perform systematically better than loans that are not sold off. This would happen if the former types of loans are characterized by high ϕ and low κ, while the latter ones have low ϕ but high κ. Therefore, a regression with a measure of loan performance as the dependent variable and a measure of selloff (i.e., whether the entire loan is sold in finite time) as an independent variable, as in Blickle et al. (2022), may yield that loans which are sold in finite time perform better. This result would, for instance, arise if the cross-sectional correlation between ϕ and κ is strongly negative and, as a consequence, that the loans sold by originators are characterized by high screening (i.e., low κ) and low monitoring (i.e., high ϕ).

We now provide a comparison of our model to alternatives in how they can generate retention and loan sales dynamics that are qualitatively similar to those observed in the data. Note that our model is able to generate retention and loan sales dynamics that are qualitatively similar to those observed in the data only when we model both screening and monitoring. Indeed, as discussed above, when there is no monitoring task—as in, e.g., Hartman-Glaser et al. (2012)—the lender retains the entire loan up to a time τ^0 and then sells its entire stake. In this case, retention is either zero or one, which is at odds with the evidence on the market for syndicated corporate loans. When there is no screening, the implementation stipulates a constant retention level and no loan sales after origination, a

pattern that is also inconsistent with the empirical evidence.

Likewise, existing dynamic asymmetric information models of asset trade—such as Daley and Green (2012) or Adelino, Gerardi, and Hartman-Glaser (2019)—feature lumpy sales, that is, the seller (the analog of the lender in our model) either holds the asset to be sold or sells it entirely and there is no partial retention. More recently, Gottardi, Moreira, and Fuchs (2022) develop a dynamic model of adverse selection in which privately informed sellers decide on how much to sell/retain of an asset when trades can take place continuously over time. They show that delay of trade dominates fractional trade as a device to achieve separation, so that in equilibrium each type trades all of its assets at a unique point in time.[26]

Finally, we would like to highlight that empirical evidence points toward moral hazard as an important driver of loan sale dynamics and vice versa. Gustafson et al. (2021) document that the extent of active monitoring crucially depends on the lead arranger's retained share and loan sales. Chen et al. (2023) and Haque et al. (2023) empirically show that changes in the severity of the lender's moral hazard problem shape loan sales.

3.3 Loan Characteristics 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 four panels of Figure 4 plot the lender's retention level β_t for $t = 0$ (solid black line), $t = 3$ (dotted red line), and $t \to \infty$ (dashed yellow line) against κ, ϕ, Λ, and γ. The lower four panels of Figure 4 plot a measure of the selloff speed, $1 - \beta_t/\beta_0$, against κ, ϕ, Λ, and γ. Note 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 initial stake quickly (slowly).

[26] While delay of trade always weakly dominates fractional trade as signaling device, this relationship is strict only under limited commitment in their setup.

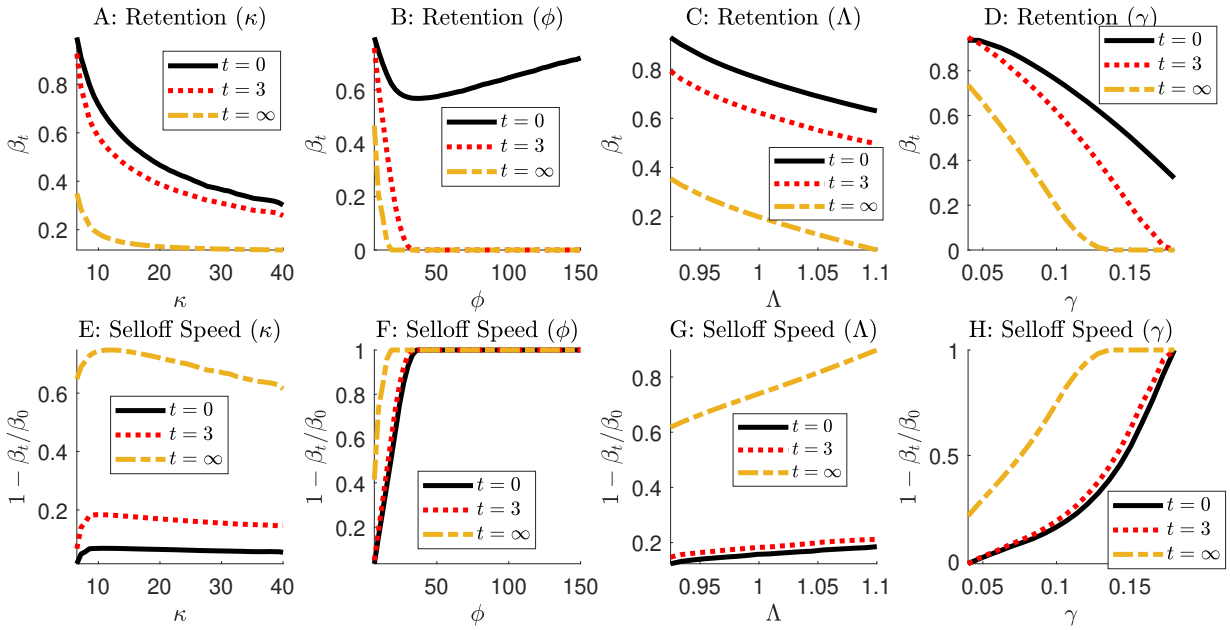

Figure 4: **Retention and dynamics**. Initial retention and selloff speed as functions of the costs of screening and monitoring κ and ϕ, intrinsic credit quality Λ, and the lender's cost of capital γ.

Figure 4 reveals that retention decreases and selloff speed increases as intrinsic credit risk Λ or the lender's discount rate γ increase (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, thus, predicts that the 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, the finding in Irani and Meisenzahl (2017) and Irani et al. (2021) that less-capitalized banks reduce loan retention, and the finding in Adelino et al. (2019) that mortgage quality is positively related to the time to sale for securitized mortgages.

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.[27] As κ increases, contracted screening and monitoring efforts decrease (see Figure 2), 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

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

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, when the cost of monitoring ϕ is sufficiently large, 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.

These results have important implications for empirical research on incentives and loan performance. Indeed, 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, while initial retention proxies for screening incentives and effort, it does not proxy monitoring incentives and effort. 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, while stimulating screening, may come along with low monitoring incentives when the originator quickly sells off its share. 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. 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.

4 Extensions and Model Variants

4.1 Loan Portfolios

Loan originators often hold a portfolio of loans. In this section, we investigate whether there are advantages in structuring lender compensation based on the performance of the overall portfolio by relaxing the loan-level limited liability. To do so, we consider two identical and independent loans $i = 1, 2$ that require separate screening and monitoring. Each loan i pays coupons at rate 1 up to its time of default τ^i. Each loan i defaults with the time-varying intensity

$$\lambda_t^i = \Lambda - q^i - a_t^i,$$

where q^i is the lender's screening of loan i at time $t = 0^-$ and a_t^i is the lender's monitoring of loan i at time t. The two loans' random default times are independent, conditional on the lender-investor contract $\mathcal{C}$. The detailed description and solution of this model variant can be found in Appendix B.2.

One possibility to incentivize loan origination for two identical loans is to write separate contracts for each loan with the lender. In this case, the baseline contract applies to each individual loan and—in the proposed contract implementation—the lender retains a time-decreasing share of each loan. The performance of one loan does not affect the value of the lender's stake in the other loan. For instance, if loan $i = 2$ defaults, the agent's stake in this loan becomes worthless, but the value of its stake in loan $i = 1$ is not directly affected. The lender is in effect protected by loan-level limited liability, i.e., the punishment the lender

incurs upon default of loan i is no larger than the loss of her stake in loan i.

We show in Appendix B.2 that such an arrangement is generally not optimal. Instead, it is optimal to relax loan-level limited liability and replace it with portfolio-level limited liability, in that the agent loses its entire stake—instead of only its stake in loan i—upon default of loan i. Structuring the lender's compensation on the portfolio level facilitates a more efficient incentive provision for screening and monitoring. As we show in Appendix B.2, the optimal contract for loan portfolios leads then to higher screening and monitoring, which reduces default risk and increases total surplus from origination.

We also propose an intuitive and practically relevant implementation of the optimal contract. In this implementation, the loan portfolio is divided into different tranches, namely, a junior/equity tranche and a senior tranche. The junior tranche is riskier and fully wiped out upon the first default event, while the senior tranche maintains its value past the first default event and absorbs only the second default event. The lender is provided incentives by retaining the junior tranche of the loan portfolio—an outcome empirically observed in the mortgage loans market (see, e.g., Begley and Purnanandam (2016))—while investors hold the senior tranche. As a result, the value of the lender's stake drops drastically if one loan defaults which in turn provides the lender incentives to screen and monitor.

4.2 The Effects of Credit Ratings and CLOs

Many loans are rated before they are sold to investors. For instance, in the market for syndicated loans, institutional investors (e.g., CLOs or loan market mutual funds) typically buy Term B loans which are most of the time rated. We now analyze how credit ratings affect the lender's incentives, retention, and loan sale dynamics. A key finding of this section is that for rated loans (e.g., Term B loans sold to institutional investors), the lender retains less of the loan and may sell its entire share shortly after origination.

In this section, we assume that with a credit rating at origination, screening effort becomes publicly observable and contractible, which removes the moral hazard over screening at origination. A micro-foundation of this assumption is provided in Appendix B.6. The intuition underlying this assumption is that the credit rating at origination reveals loan quality

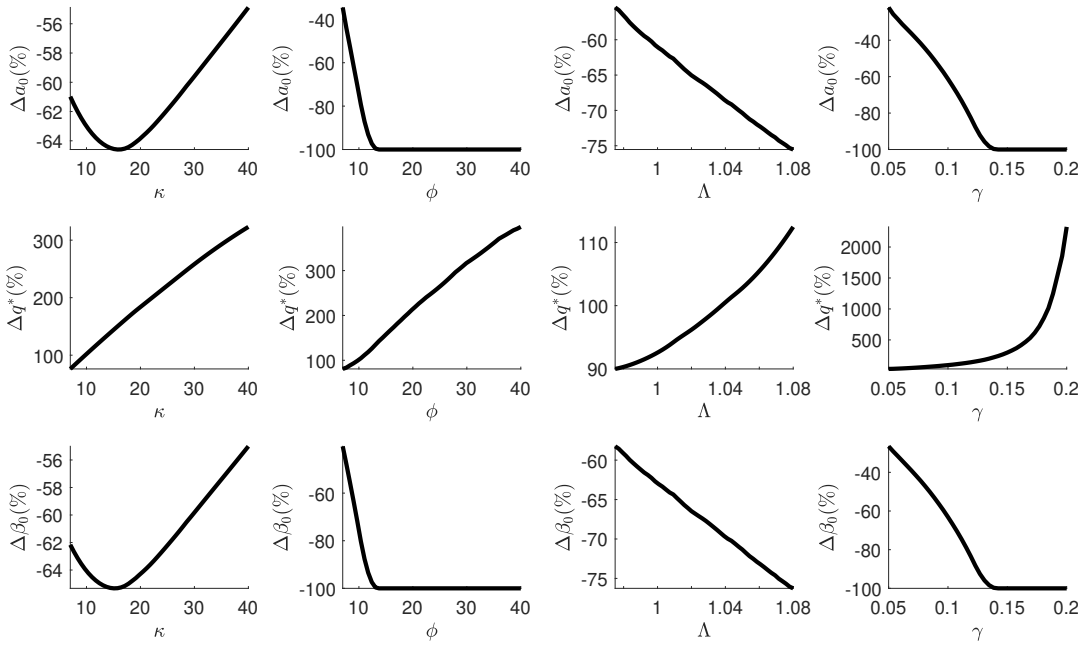

Figure 5: **The effects of credit ratings.** Δy denotes the percentage change in the initial value of the outcome variable y caused by a credit rating, where $y \in \{a_0, q^*, \beta_0\}$. Outcome variables are plotted as functions of the cost of monitoring κ, the cost of screening ϕ, the raw default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

and 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. The credit rating increases screening at origination but reduces monitoring a_0. The reason is that the credit rating increases the lender's incentives to screen loans at origination without requiring increasing its skin in the game. The lender, therefore, retains a lower share in the loan β_0, leading to lower monitoring incentives $W_0 = \phi a_0$. In the market for syndicated loans, Term B loans are typically rated and sold to institutional investors, such as CLOs or loan market mutual

funds. Our findings on the effects of credit ratings imply that such loans are subject to more screening at origination and less monitoring after origination. Finally, our model predicts that the share retained by the originator should be lower when the originator sells rated loans to CLOs. The bottom row in Figure 5 indicates that the retention of rated loans is particularly low when ϕ or γ are large.

4.3 The Effects of Loan Maturity

Our baseline model features infinite maturity loans or finite maturity loans that are rolled over up to default. We now consider finite maturity loans that are not rolled over. The extension is important for two reasons. First, we want to show that our main results do not hinge on a specific modeling of maturity. Second, as screening and monitoring efforts have effects of different duration, loan maturity—which affects a loan's duration—could have different effects on these two tasks, with important implications for how loan maturity shapes the lender's dynamic retention and loan sales.

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, the firm pays back the face value, which is the joint terminal payoff of the lender and outside investors. The baseline setting corresponds to the case $\delta = 0$.

The model with a finite maturity and its solution are described in Appendix B.3. 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 state variables. The contract dynamics in the model with finite maturity are qualitatively similar to those in the baseline model, and the contract can be implemented by requiring the originator to hold a time-decreasing share of the loan. As we show in Appendix B.3, the agent's screening incentives at time $t = 0$ read

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

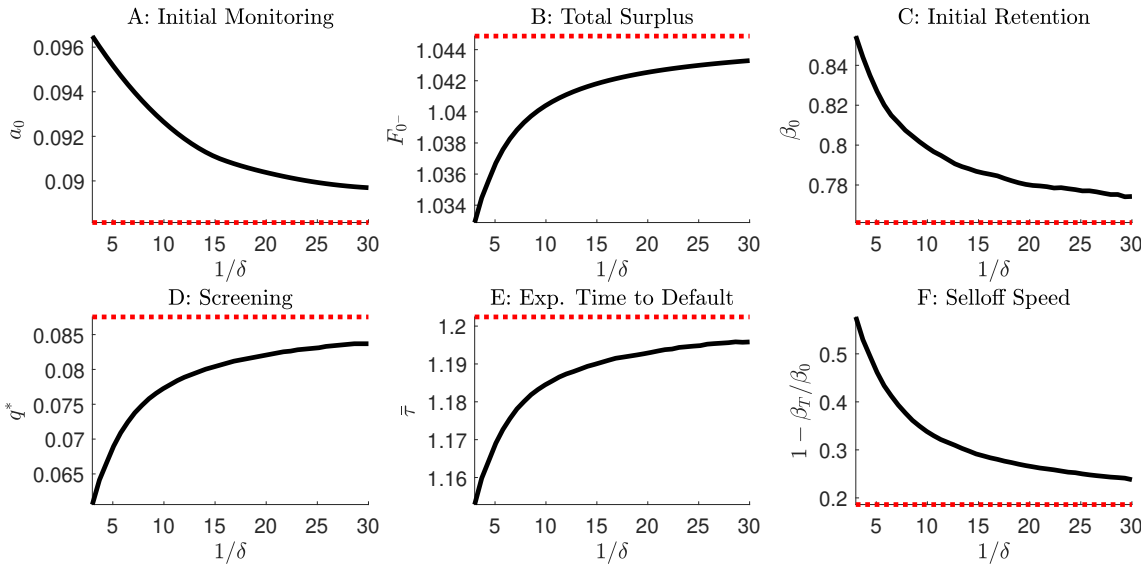

Figure 6: **The effects of debt maturity**. We use our baseline parameters and set $T = 3$ for selloff speed. The dotted red line depicts the outcomes with infinite debt maturity.

which is the product of the value and the duration of the lender's exposure. At maturity, the lender exits and is no longer exposed to default risk, so its screening incentives fall to zero; thus, the difference between (10) and (31) is that δ augments the discount rate, which reduces screening incentives. That is, keeping the value of the lender's claim constant, a shorter maturity reduces the duration of the claim and thus the lender's long-run exposure to loan performance, thereby undermining screening incentives. In contrast, loan maturity has no direct effect on monitoring incentives, as the impact of monitoring is short-lived.

The total effect of finite maturity also depends on its impact on the value of the lender's claim. Figure 6 plots initial monitoring effort a_0 (Panel A) and screening effort q^* (Panel D) for varying loan maturities. Short maturity undermines screening incentives by shortening the duration of the lender's claim. To counteract this adverse effect, the optimal contract stipulates a higher value of the lender's initial exposure W_0 which leads to high monitoring effort a_0 for short maturity loans (Panel A). Despite high initial exposure, the duration effect dominates, and so screening effort decreases for short maturity loans (Panel D). Therefore, our model predicts relatively low screening but high initial monitoring for loans with a short maturity. Implementing these incentives for short maturity loans requires a higher initial

retention level β_0 (Panel C) and a relatively quick selloff (Panel D) after origination.

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.[28] Panel B of Figure 6 shows that total surplus increases with debt maturity due to lower agency costs.

4.4 Repeated Loan Origination

Repeated lender-borrower and lender-investor interactions are common in credit markets, in particular in syndicated lending. To formally analyze repeated lender-investor interactions, we consider that (after screening) the lender originates a loan with face value K, coupon payments at rate 1, and stochastic maturity arriving with exogenous intensity $\delta > 0$ at time τ^δ. When the loan matures at time τ^δ, the face value is repaid, and the lender originates a new and identical loan to a new borrower with exogenous probability p^δ, so that the lender has to re-screen and exert costly screening effort when a new loan is originated.[29] Otherwise, with probability $1 - p^\delta$, the lender exits and cannot originate further loans. If a loan defaults at time τ, the lender originates another loan with (exogenous) probability p^λ. With probability $1 - p^\lambda$, the lender exits (i.e., the relationship breaks down). We introduce the probabilities p^δ and p^λ to make the continuation of the relationship probabilistic, thereby capturing the fact that the number of repeated interactions is not infinite in practice. For simplicity, we assume that $p^\delta = 1$ and $p^\lambda \leq 1$. That is, the lender faces the possibility of being excluded from the secondary market upon default and not being able to originate further loans (due, e.g., to a loss in competitiveness related to the inability to sell the loan).

[28] To 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

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

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

[29] Note that our baseline setting already captures repeated lender-borrower interactions. Notably, our baseline model can be interpreted as follows: The lender extends a loan with face value K and coupon payments at rate 1 to the borrower with a possibly finite maturity, and this loan is rolled over at identical terms and without re-screening at maturity until a default occurs at time τ.

This assumption is consistent with empirical evidence in Gopalan et al. (2011) showing that, following poor performance (e.g., a default of a previously syndicated loan), lead arrangers sell less of the loans they syndicate. Setting $p^{\lambda} = 0$ and $\delta = 0$ yields our baseline model.

A detailed description of this model variant as well as its solution and analysis are presented in Appendix B.4. In this Appendix, we derive two main findings. First, the going-concern value from repeated loan origination and sale serves as an incentive mechanism for screening and monitoring, which substitutes for loan retention (as found in Gopalan et al. (2011)).[30] Thus, with repeated interactions, the lender generally retains a lower share or sells the loan faster. As such, due to repeated origination and lender-investor relationships, the lender may have strong incentives to screen and monitor so that the loan's default risk is low, even if loan retention by the originator is low and or if it sells its share quickly after origination. This provides a rationale for the finding that, in the market for syndicated loans, loans need not perform poorly even when the lead lender sells its share relatively quickly after origination (Blickle et al., 2022).

Second, repeated interactions facilitate lender commitment to a specific retention path stipulated in the contract implementation. Recall from Section 3.1 that the optimal contract can be implemented by having the lender retain a time-decreasing share of the loan, thus inducing optimal loan sales as part of the optimal full-commitment contract. This implementation also applies in the context of the model variant with repeated lender-investor interactions. Because the lender has a higher cost of capital and thus values the loan less than investors, the lender might be tempted to deviate from the recommended retention schedule and sell a larger share of the loan to investors. In Appendix B.4, we consider that if the lender were to deviate by selling more than recommended, investors would cut the relationship, thereby precluding further loan origination by the lender. That is, investors play a grim-trigger strategy. We then show that, provided the lender expects sufficiently large payoffs from future loan origination, the lender indeed finds it optimal not to deviate from the contracted retention path, even when it cannot commit.

[30] Hartman-Glaser (2017) obtains a similar result in a model of repeated asset sales under adverse selection, showing that reputation concerns can substitute for retention in signaling.

4.5 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 servicing 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 is in charge of monitoring. The 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 B.5 provides a detailed description and solution of 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.

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 (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, as we show in Appendix B.5, 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). 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). The analysis also predicts that 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 and syndicated loans.

5 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 (2023) 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. (2023). Busy venture capitalists and investment performance. *Working paper, The University of Virginia*.

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.

Begley, T. A. and A. Purnanandam (2016). Design of financial securities: Empirical evidence from private-label rmbs deals. *Review of Financial Studies 30*(1), 120–161.

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*.

Blickle, K., C. Parlatore, and A. Saunders (2023). Specialization in banking. Technical report, National Bureau of Economic Research.

Board, S. and M. Meyer-ter Vehn (2013). Reputation for quality. *Econometrica 81*(6), 2381–2462.

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.

Bruche, M., F. Malherbe, and R. R. Meisenzahl (2020). Pipeline risk in leveraged loan syndication. *The Review of Financial Studies 33*(12), 5660–5705.

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.

Chen, M., S. J. Lee, D. Neuhann, and F. Saidi (2023). Less bank regulation, more non-bank lending. *Working Paper*.

Daley, B. and B. Green (2012). Waiting for news in the market for lemons. *Econometrica 80*(4), 1433–1504.

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

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

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

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

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

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

Gopalan, R., V. Nanda, and V. Yerramilli (2011). Does poor performance damage the reputation of financial intermediaries? evidence from the loan syndication market. *The Journal of Finance 66*(6), 2083–2120.

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

Gottardi, P., H. Moreira, and W. Fuchs (2022). Time trumps quantity in the market for lemons. Technical report.

Gryglewicz, S. and S. Mayer (2022). Dynamic contracting with intermediation: Operational, governance, and financial engineering. *Journal of Finance (forthcoming)*.

Gryglewicz, S., S. Mayer, and E. Morellec (2021). Optimal financing with tokens. *Journal of Financial Economics 142*(3), 1038–1067.

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.

Haque, S., S. Mayer, and T. Wang (2023). How private equity fuels non-bank lending. *Working Paper*.

Hartman-Glaser, B. (2017). Reputation and signaling in asset sales. *Journal of Financial Economics 125*(2), 245–265.

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

Heitz, A. R., C. Martin, and A. Ufier (2022). Bank monitoring with on-site inspections. *FDIC Center for Financial Research Paper No. 2022-09*.

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.

Hoffmann, F. and S. Pfeil (2021). Dynamic multitasking and managerial investment incentives. *Journal of Financial Economics 142*(2), 954–974.

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

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

Hu, Y. and F. Varas (2021b). A theory of zombie lending. *The Journal of Finance 76*(4), 1813–1867.

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.

Irani, R. M. and R. R. Meisenzahl (2017). Loan sales and bank liquidity management: Evidence from a us credit register. *The Review of Financial Studies 30*(10), 3455–3501.

Ivashina, V. (2009). Asymmetric information effects on loan spreads. *Journal of Financial Economics 92*(2), 300–319.

Jiang, S., S. Kundu, and D. Xu (2023). Monitoring with small stakes. *Available at SSRN 4271851*.

Keys, B. J., T. Mukherjee, A. Seru, and V. Vig (2010). Did securitization lead to lax screening? evidence from subprime loans. *The Quarterly journal of economics 125*(1), 307–362.

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.

Mayer, S. (2022). Financing breakthroughs under failure risk. *Journal of Financial Economics 144*(3), 807–848.

Nadauld, T. D. and M. S. Weisbach (2012). Did securitization affect the cost of corporate debt? *Journal of financial economics 105*(2), 332–352.

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.

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

Purnanandam, A. (2011). Originate-to-distribute model and the subprime mortgage crisis. *The review of financial studies 24*(6), 1881–1915.

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 Proofs

A.1 Proof of Lemma 1

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

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

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 (6), 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 in an incentive compatible contract satisfies $\hat{a}_t = W_t/\phi$.

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

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

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 (10) 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

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

is the unique solution to (A.2) if

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

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 (9), 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 (12) to obtain

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

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

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

for all $t \geq 0$. In addition, (10) implies

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

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

$$\lambda_t = \Lambda - a_t - q \geq \Lambda - \bar{a} - \bar{q}. \tag{A.8}$$

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

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

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

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

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

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

That is, when (A.11) holds, the first order approach is valid and (A.3) or, equivalenty, (9) (due to $q < \bar{q}$) pins down screening effort. Note that (A.11) is equivalent to condition (13) (Lemma 1). Also notice that (13) but not per-se necessary.

A.2 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

$$dP_t = (r+\lambda_t)P_t dt - 1dt + dC_t \tag{A.12}$$

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

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

$$\begin{aligned} dF_t &= dP_t + dW_t && \text{(A.14)} \\ &= (r+\lambda_t)P_t dt - 1dt + dC_t + (\gamma+\lambda_t)W_t dt - dC_t + \frac{\phi a_t^2}{2}dt \\ &= (r+\lambda_t)(\underbrace{P_t + W_t}_{=F_t})dt - 1dt + \frac{\phi a_t^2}{2}dt - (\gamma - r)W_t dt. && \text{(A.15)} \end{aligned}$$

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

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

which is (20) from the main text.

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

$$F_0 - \frac{\kappa q^2}{2}, \tag{A.17}$$

where F_0 is characterized in (A.16). Note that it is always possible to stipulate payouts dC_t to the agent, which decreases W_t by some 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 (6) 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

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

subject to the monitoring incentive constraint (6), which can be rewritten as (17).

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

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

and the optimal lender continuation value is $W^B(q) = \phi a^B(q)$, due to (6). 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 (A.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., $\frac{\partial}{\partial q} a^B(q) \geq 0$.

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

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

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

The objective (A.17) can be rewritten as

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

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

A.3 Proof of Proposition 2

A.3.1 Preliminaries

To begin, we derive the dynamics of W_t, i.e., (11), the dynamics of V_t (defined in (8)), and the integral expression (10). Now, recall the definition of W_t in (4) and differentiate (4) 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 (11). Using (11), we can write the intermediary's optimization with respect to monitoring effort a_t at time t as

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

which yields optimal $a_t=\min\left\{\frac{W_t}{\phi},\bar{a}\right\}$ (as in (6)) 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 (A.22) under optimal a_t with respect to q so that

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

which is (12) 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 $\frac{\partial c_t}{\partial q}=0$.[31] We can integrate $\dot{V}_t=(\gamma+\lambda_t)V_t-W_t$ over time t

[31] In 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 (11) by inserting monitoring incentive compatibility, $a_t=W_t/\phi$, to obtain

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

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

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

to obtain the integral expression (10), 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 function $F(V)$) solves the HJB equation (23). Part II demonstrates that $\lim_{t\to\infty} V_t = V^B(q)$. Part III characterizes the agent's initial choice of optimal screening effort $q = q^*$. 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. Unless otherwise mentioned, we focus 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 (23) 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).

A.3.2 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 (9) holds at time $t = 0$ so that $V_0 = \kappa q$.

The optimal contract maximizes total surplus characterized in (A.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 some 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 (12), 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

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

which is solved subject to the monitoring incentive constraint (6). Recall that both the principal and the agent are subject to limited liability, so that $W \in [0, F(V)]$ and the principal's payoff $F(V) - W$ satisfies $F(V) - W \in [0, F(V)]$ too. The above HJB equation coincides with (23). The

which simplifies to (12), as desired.

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

$$a(V) = \frac{F(V) - F'(V)(V+\phi) - (\gamma - r)\phi}{\phi} \quad \wedge \quad \frac{F(V)}{\phi}. \tag{A.23}$$

With slight abuse of notation, when above expression is negative, then $a(V) = 0$. 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 (12), we can solve for

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

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

$$F(V^B(q)) = F^B(q). \tag{A.25}$$

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 $\lim_{t\to\infty} V_t = V^B(q)$, which then—together with (A.25)—implies

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

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

A.3.3 Part II

In this part, we prove that $\lim_{t\to\infty} V_t = V^B(q)$. To do so, we set up the Lagrangian for the total surplus maximization at time $t = 0$

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

where ℓ is the Lagrange multiplier with respect to the screening incentive constraint (9) and $W_t = \phi a_t$ is the effort incentive constraint which we directly insert into the objective function.

Next, we rewrite (A.14) as

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

which can be integrated over time to obtain

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

Likewise, we can rewrite (12) as

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

which can be integrated over time to get

$$V_t = \int_t^\infty e^{-\gamma(s-t)} \left(W_s - \lambda_s V_s\right) ds. \tag{A.28}$$

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

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

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 (6), i.e., $a_t = W_t/\phi$. If interior, optimal effort a_t satisfies the first order condition:

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

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

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

Accounting for limited liability $W_t = \phi a_t \leq F_t$ and $a_t \geq 0$, we can solve (A.31) for

$$a_t = \max\left\{0, \frac{F_t - (\gamma - r)\phi - \ell e^{-(\gamma - r)t}(V_t + \phi)}{\phi}\right\} \quad \wedge \quad \frac{F_t}{\phi}. \tag{A.32}$$

Taking the limit $t \to \infty$ in (A.32) upon noticing that $\lim_{t\to\infty} W_t < \lim_{t\to\infty} F_t$ leads to

$$\lim_{t\to\infty} a_t = \lim_{t\to\infty} \left(\frac{F_t - (\gamma - r)\phi}{\phi}\right) < \lim_{t\to\infty} \frac{F_t}{\phi}, \tag{A.33}$$

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 $t \to \infty$, the solution becomes stationary and F_t and a_t become constant, in that

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

for (endogenous) constants $\hat{F}$ and $\hat{a}$.[32] Note that by (A.33),

$$\hat{a} = \max\left\{0, \frac{\hat{F} - (\gamma - r)\phi}{\phi}\right\}. \tag{A.34}$$

[32] Equivalently,

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

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

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

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

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

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

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

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

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 (A.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 (A.37). Likewise, when $V_0 = \kappa q < V^B(q)$, it follows that $\dot{V}_t > 0$, with convergence according to (A.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 $\lim_{V\to V^B(q)} F'(V)$. Note that due to (A.25), that is, $F(V^B(q)) = F^B(q)$, and $\lim_{t\to\infty} V_t = V^B(q)$, it follows that $\lim_{V\to V^B(q)} F(V) = F^B(q)$ and $\lim_{V\to V^B(q)} a(V) = a^B(q)$. We know from Proposition 1 that $W^B(q) < F^B(q)$, so that $\lim_{V\to V^B(q)} W(V) < \lim_{V\to V^B(q)} F(V)$. Thus, for V close to $V^B(q)$, the principal's limited liability constraint does not bind. Using (A.23), $\lim_{V\to V^B(q)} a(V) = a^B(q)$ becomes equivalent to

$$\lim_{V\to V^B(q)} F'(V) = 0, \tag{A.38}$$

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

$$\lim_{V\to V^B(q)} F'(V) = \frac{F^B(q) - (\gamma - r)\phi}{\phi} \leq 0, \tag{A.39}$$

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

A.3.4 Part III

At time $t = 0$, initial screening incentive V_0 pins down screening effort q by means of the screening incentive constraint (9). The agent picks the amount of initial screening incentives V_0 to maximize

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

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 (9) is optimally tight.

The first order condition to (A.40) is

$$\frac{\partial F(V_0)}{\partial q}|_{q=q^*} + F'(V_0)\kappa = \kappa q^*, \tag{A.41}$$

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

A.3.5 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 $\lim_{t\to\infty} V_t = V^B(q^*)$ and because there is no stationary point with $\dot{V}_t = 0$, $V_0 > V^B(q^*)$ implies $\dot{V}_t < 0$ whenever $V_t > V^B(q^*)$. It suffices to consider $q^* > 0$ and $a^B(q^*) > 0$.

Suppose to the contrary that

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

where the last equality follows (A.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

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

where the first inequality follows (A.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, (A.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 (A.18) with respect to q to obtain under the optimal controls $(W^B(q), a^B(q))$:

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

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.[33] Using (A.44), we obtain

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

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:

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

with $V^B(q)$ from (A.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

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

[33] To see this, note that $\frac{\partial a^B(q)}{\partial q} = \frac{\partial F^B(q)}{\partial q} \frac{1}{\phi}$. Thus, differentiating (A.44) with respect to q:

$$(r + \lambda) \frac{\partial^2 F^B(q)}{\partial q^2} = \frac{\partial F^B(q)}{\partial q} + \frac{1}{\phi} \left(\frac{\partial F^B(q)}{\partial q} \right)^2 > 0.$$

Differentiating this relationship with respect to q:

$$(r + \lambda) \frac{\partial^3 F^B(q)}{\partial q^3} = \frac{\partial^2 F^B(q)}{\partial q^2} + \frac{2}{\phi} \frac{\partial F^B(q)}{\partial q} \frac{\partial^2 F^B(q)}{\partial q^2} + \frac{\partial a^B(q)}{\partial q} \frac{\partial^2 F^B(q)}{\partial q^2} > 0.$$

where the first inequality uses that $r < \gamma$ and $W^B(q) \leq F^B(q)$ and the last equality uses (A.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 (A.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 (6) 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 (14). 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

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

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

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

which — by (A.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, $\lim_{V \to V^B(q^\varepsilon)} F'(V) = 0$ (see (A.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 $q = q^*$.

A.3.6 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 (A.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 (23) 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 and $\dot{V} \neq 0$, we can differentiate above ODE with respect to V to calculate

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

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 $V > V^B(q)$ and $\dot{V} \neq 0$. 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 (23) under the optimal controls with respect to V, which yields

$$F''(V) = \frac{-(\gamma - r)F'(V)}{(\gamma + \lambda)V - W}. \tag{A.49}$$

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 (A.49) 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 $V \to V^B(q)$). When $a^B(q) = 0$, then (A.39) implies $\lim_{V \to V^B(q)} F'(V) \leq 0$. Otherwise, when $a^B(q) > 0$, then (A.38) implies $F'(V^B(q)) = 0$ and — according to the expression for effort (A.23):

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

owing to $\gamma > r$.

If it were $F'(V), F''(V) \geq 0$ in a right-neighborhood 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

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

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 $(V^B(q), V_0)$.

A.3.7 Part VI

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

We can solve (11) to get the payout rate

$$c_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \dot{W}_t. \tag{A.50}$$

If $F_t = W_t$, note that according to (A.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 (A.50) yields $c_t = 1 > 0$. Further, provided $a(V)$ is differentiable, we have $a'(V) = F'(V)/\phi < 0$, so that $\dot{a}_t = a'(V_t)\dot{V}_t > 0$.

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

$$a(V) = \max\left\{0, \frac{F(V) - F'(V)[V + \phi] - (\gamma - r)\phi}{\phi}\right\},$$

and, provided $a(V)$ is differentiable, then $a'(V) = \frac{-F''(V)[V+\phi]}{\phi} > 0$, as $F''(V) < 0$. Thus, $\dot{a}_t = a'(V_t)\dot{V}_t < 0$ and, by (6), $\dot{W}_t < 0$. Inserting $\dot{W}_t < 0$ into (A.50) implies $c_t > 0$.

A.4 Proof of Proposition 3 and details on the implementation

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

Next, we provide more details for the implementation and 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

$$\lim_{V \to V^B(q)} L'(V) = 0 \quad \Longleftrightarrow \quad \lim_{V \to V^B(q)} L(V) = \frac{1}{r + \Lambda - a^B(q) - q},$$

whereby $\lim_{V \to V^B(q)} \dot{V} = 0$ and $\lim_{V \to v^B(q)} a(V) = a^B(q)$.

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

$$\beta(V) - \beta'(V)\dot{V}L(V) = (\gamma + \Lambda - a(V) - q)W(V) + \frac{\phi a(V)^2}{2} - W'(V)\dot{V}, \tag{A.51}$$

which is solved subject to

$$\lim_{V \to V^B(q)} \beta'(V) = 0 \iff \lim_{V \to V^B(q)} \beta(V) = c^B(q) = (\gamma + \Lambda - a^B(q) - q)W^B(q) + \frac{\phi(a^B(q))^2}{2}. \tag{A.52}$$

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 (A.51), as desired. Throughout, we assume the existence and uniqueness of a (non-constant) solution to (A.51) subject to (A.52) on $(V^B(q), V_0]$.

Finally, we show that $L_t(1 - \beta_t) = P_t = F_t - W_t$. For this sake, take

$$P_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du}(1 - c_s)ds = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du}(1 - \beta_s + \dot{\beta}_s L_s)ds$$

so that

$$\dot{P}_t = (r + \lambda_t)P_t - (1 - \beta_t + \dot{\beta}_t L_t).$$

Next calculate

$$\dot{L}_t = (r + \lambda_t)L_t - 1.$$

We start by conjecturing that $P_t = (1 - \beta_t)L_t$, and in what follows verift this conjecture. We calculate

$$\begin{aligned}\dot{P}_t &= (r + \lambda_t)P_t - (1 - \beta_t + \dot{\beta}_t L_t) = (r + \lambda_t)(1 - \beta_t)L_t - (1 - \beta_t + \dot{\beta}_t L_t) \\ &= (1 - \beta_t)\dot{L}_t - \dot{\beta}_t L_t = \frac{d}{dt}\big[(1 - \beta_t)L_t\big],\end{aligned}$$

where the second equality uses $P_t = (1 - \beta_t)L_t$, the third equality uses $\dot{L}_t + 1 = (r + \lambda_t)L_t$ and simplifies, and the fourth equality collects terms. Thus, $P_t = (1 - \beta_t)L_t$ implies $\dot{P}_t = \frac{d}{dt}\big[(1 - \beta_t)L_t\big]$.

To conclude our argument we consider the limit $t \to \infty$ in which case $V_t \to V^B(q)$ as well as $W_t \to W^B(q)$, $F_t \to F^B(q)$, and $L_t \to L^B(q) = \frac{1}{r + \lambda^B(q)}$. Then, under the optimal controls $a^B(q), W^B(q)$ and payouts $c^B(q) = (\gamma + \lambda^B(q))W^B(q) + \frac{\phi(a^B(q))^2}{2} = \beta^B(q)$, we have

$$P^B(q) = \frac{1 - c^B(q)}{r + \lambda^B(q)} = \frac{(1 - \beta^B(q))}{r + \lambda^B(q)} = (1 - \beta^B(q))L^B(q).$$

Thus, in the limit $t \to \infty$, we have $P_t \to P^B(q)$ as well as $(1 - \beta_t)L_t \to P^B(q)$, i.e., $\lim_{t\to\infty} P_t = \lim_{t\to\infty}(1 - \beta_t)L_t$. Because, in addition, $P_t = (1 - \beta_t)L_t$ implies $\dot{P}_t = \frac{d}{dt}\big[(1 - \beta_t)L_t\big]$ holds, we have $P_t = (1 - \beta_t)L_t$ at all times.

A.5 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.

We now prove the second claim about the limit case of $\phi \to \infty$. For this sake, fix q. The below arguments hold for any q, including the optimal $q = q^*$ determined at time $t = 0^-$. For given V and q, we use the notation $\hat{x} = \lim_{\phi\to\infty} x$.

To begin, recall $\lim_{V\to V^B(q)} F(V) = F^B(q)$, $\lim_{V\to V^B(q)} a(V) = a^B(q)$, and $\lim_{V\to V^B(q)} W(V) = W^B(q)$ with

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

$$a^B(q) = \max\left\{\frac{F^B(q) - (\gamma - r)\phi}{\phi}, 0\right\},$$

$$W^B(q) = \phi a^B(q) = \max\left\{F^B(q) - (\gamma - r)\phi, 0\right\},$$

and

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

Because $F^B(q)$ is bounded (specifically, $F^B(q) < \frac{1}{r+\Lambda-\bar{a}-\bar{q}}$), it is clear that there exists $\phi' > 0$ such that for all $\phi > \phi'$, $a^B(q) = W^B(q) = V^B(q) = 0$. In particular, in the limit $\phi \to \infty$, we have $\hat{a}^B = 0$ as well as $\hat{W}^B = 0$ and $\hat{V}^B(q) = 0$. Thus, the relevant interval for the state variable V, $(V^B(q), V_0]$, becomes $(0, V_0]$. We restrict attention to levels of V lying in this interval.

Next, recall from (A.23) that the optimal effort $a(V)$ solves

$$a(V) = \frac{F(V) - F'(V)(V + \phi) - (\gamma - r)\phi}{\phi} \quad \wedge \quad \frac{F(V)}{\phi}, \tag{A.53}$$

Because of $\lim_{V\to 0} a(V) = a^B(q)$ with $a(V) > 0$ in a right-neighbourbood of zero, we have

$$\lim_{V\to 0} F'(V) = \begin{cases} 0 & \text{if} \quad a^B(q) > 0 \\ \frac{F^B(q) - (\gamma - r)\phi}{\phi} & \text{if} \quad a^B(q) = 0. \end{cases}$$

As argued above, there exists $\phi' > 0$ such that for all $\phi > \phi'$, $a^B(q) = 0$ and therefore $\lim_{V\to 0} F'(V) = \frac{F^B(q) - (\gamma - r)\phi}{\phi}$. Because the value function is strictly concave, we have

$$F'(V) < \frac{F^B(q) - (\gamma - r)\phi}{\phi}$$

for all $V > 0$. We assume that for any $V \in (0, \kappa q]$, the limit $\lim_{\phi\to\infty} F(V) = \hat{F}(V)$ exists and that the function $\hat{F}(V)$ is twice continuously differentiable and strictly concave, i.e., $\hat{F}''(V) < 0$.

Next, for $V > 0$, we can take the limit $\phi \to \infty$ to obtain:

$$\hat{F}'(V) \leq \lim_{\phi\to\infty} \left(\frac{F^B(q) - (\gamma - r)\phi}{\phi}\right) = -(\gamma - r).$$

Due to strict concavity of $\hat{F}(V)$, i.e., $\hat{F}''(V) < 0$ for $V > 0$, it follows that above inequality is strict, i.e., $\hat{F}'(V) < -(\gamma - r)$ for $V > 0$.

Next, using (A.53), we have

$$W(V) = \phi a(V) = \min\{F(V) - F'(V)(V + \phi) - (\gamma - r)\phi, F(V)\}. \tag{A.54}$$

We can take the limit $\phi \to \infty$ for optimal continuation payoff in (A.54), which, conditional on $\hat{F}'(V) < -(\gamma - r)$, is $\hat{W}(V) = \hat{F}(V)$. It follows

$$\lim_{V\to 0}\lim_{\phi\to\infty} W(V) = \lim_{V\to 0} \hat{W}(V) = \lim_{V\to 0} \hat{F}(V) > \hat{W}(0) = \lim_{\phi\to\infty}\lim_{V\to 0} W(V) = \lim_{\phi\to\infty} W^B = 0.$$

As $\hat{W}(V)$ is dis-continuous and exhibits an upward jump at $V = 0$, it follows that $\hat{W}(V_t)$ drops down once V_t reaches zero (from above). Moreover,

$$\lim_{\phi\to\infty} \dot{V} = (r + \hat{\lambda}(V))V - \hat{W}(V)$$

is strictly negative in an open right-neighbourhood of $V = 0$, so that V reaches zero in finite time $\tau^0 = \inf\{t \geq 0 : V_t = 0\}$ in the limit $\phi \to \infty$.

We can rewrite the continuation payoff allowing for a general payment process

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

Thus

$$dW_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - dC_t \quad \Longleftrightarrow \quad dW(V) = (\gamma + \lambda(V))W(V) + \frac{\phi a(V)^2}{2} - dC(V).$$

In the limit,

$$d\hat{W}(V) = (\gamma + \hat{\lambda}(V))\hat{W}(V) - d\hat{C}(V).$$

It follows that at time τ^0 (once V_t reaches zero), there is a lumpy payout

$$d\hat{C}_{\tau^0} = d\hat{C}(0) = \lim_{V\to 0} \hat{W}(V) = \lim_{V\to 0} \hat{F}(V) = \frac{1}{r + \Lambda - q}.$$

Recall (20) so that $dF(V) = (r + \lambda(V))F(V)dt + \frac{\phi a(V)^2}{2}dt + (\gamma - r)W(V)dt - 1dt$. Before time τ^0, i.e., for $V > 0$, we have $\hat{F}(V) = \hat{W}(V)$, implying

$$d\hat{W}(V) = (\gamma + \hat{\lambda}(V))\hat{W}(V)dt - d\hat{C}(V) = d\hat{F}(V) = (r + \hat{\lambda}(V))\hat{F}(V) + (\gamma - r)\hat{W}(V)dt - 1dt,$$

which — due to $\hat{F}(V) = \hat{W}(V)$ implies $d\hat{C}(V) = 1dt$ for $t < \tau^0$.

The implementation of the optimal contract then satisfies

$$\hat{\beta}(V)dt - d\hat{\beta}(V)L = d\hat{C}(V),$$

where $L = 1/(r+\Lambda-q)$ is the loan's fair market value. At time τ^0, i.e., for $V = 0$, we have $d\hat{C}_t = L$

so that $d\hat{\beta}_{\tau^0} = -1$. Before time τ^0, i.e., for $V > 0$, we have $d\hat{C}(V) = 1dt$. Thus, above relationship holds for $\hat{\beta}(V) = 1$, which concludes the argument.

Finally, we verify strict concavity of $\hat{F}(V)$. For this sake, take the limit $\phi \to \infty$ in the HJB equation (23) for $V > 0$ noticing that $\hat{W}(V) = \hat{F}(V)$, $\hat{a}(V) = \phi(\hat{a}(V))^2 = 0$ to obtain

$$(r + \hat{\lambda}(V))\hat{F}(V) = 1 - (\gamma - r)\hat{W}(V) + \hat{F}'(V)((\gamma + \hat{\lambda}(V))V - \hat{W}(V)),$$

which—after inserting $\hat{W}(V) = \hat{F}(V)$—is equivalent to

$$(\gamma + \hat{\lambda}(V))\hat{F}(V) = 1 + \hat{F}'(V)((\gamma + \hat{\lambda}(V))V - \hat{F}(V)).$$

We can take the derivative with respect to V to obtain:

$$\hat{F}''(V) = \frac{(\hat{F}'(V))^2}{(\gamma + \lambda)V - F(V)} < 0.$$

A.6 Proof of Corollary 1

A.6.1 Part 1

A necessary condition for the lender's stake to approach zero is that $a^B = \lim_{t\to\infty} \beta_t = \beta^B = c^B = 0$. We first show that when ϕ is sufficiently large and satisfies the condition presented in the Proposition, it follows that $a^B = c^B = 0$. We take the (optimal) screening level $q^* = q$ as given.

First, we recall that given q:

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

Because $\frac{\partial F^B}{\partial q} > 0$, we obtain

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

i.e., an upper bound for $F^B = F^B(q)$ that does not depend on q.

Next, the first order derivative with respect to a satisfies

$$\frac{\partial(\Lambda + r)F^B}{\partial a} = F^B - (\gamma - r)\phi - \phi a \quad \Longleftrightarrow \quad \frac{\partial F^B}{\partial a} = \frac{F^B - (\gamma - r)\phi - \phi a}{\Lambda + r}. \tag{A.55}$$

We have $a = a^B = 0$ when

$$\frac{\partial F^B}{\partial a}|_{a=0} \leq 0 \quad \Longleftrightarrow \quad F^B|_{a=0} \leq (\gamma - r)\phi.$$

Owing to $F^B|_{a=0} \leq \frac{1}{r+\Lambda-\bar{q}}$, we obtain $F^B|_{a=0} \leq (\gamma - r)\phi$ if

$$\frac{1}{r + \Lambda - \bar{q}} \leq (\gamma - r)\phi \quad \Longleftrightarrow \quad \phi \geq \frac{1}{(r + \Lambda - \bar{q})(\gamma - r)}.$$

Next, notice that the second order derivative of F^B with respect to a satisfies:

$$\frac{\partial^2(\Lambda+r)F^B}{\partial a^2} = \frac{\partial F^B}{\partial a} - \phi = \frac{F^B - (\gamma-r)\phi - \phi a}{\Lambda+r} - \phi$$
$$< \frac{1}{\Lambda+r}\left(\frac{1}{r+\Lambda-\bar{q}-\bar{a}} - (\gamma-r)\phi\right) - \phi = \frac{1}{\Lambda+r}\left(\frac{1}{r+\Lambda-\bar{q}-\bar{a}} - (\gamma+\Lambda)\phi\right),$$

where the second equality uses (A.55), the inequality uses $F^B - \phi a < \frac{1}{r+\Lambda-\bar{q}-\bar{a}}$, and the last equality collects terms.

Thus, we obtain $a^B = 0$ if

$$\phi > \max\left\{\frac{1}{(r+\Lambda-\bar{q})(\gamma-r)}, \frac{1}{(r+\Lambda-\bar{q}-\bar{a})(\gamma+\Lambda)}\right\},$$

i.e., if ϕ is sufficiently large. As such, we have $a^B = V^B = W^B = 0$, $\lim_{t\to\infty} V_t = \lim_{t\to\infty} \dot{V}_t = 0$, as well as $c^B = 0$ and $\lim_{t\to\infty} c_t = 0$.

Next, note that

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

As such, a necessary condition condition for $a^B = 0$ — that is, $\frac{\partial F^B}{\partial a}|_{a=0} \leq 0$ — and thus $\beta^B = 0$ is that (see (A.55))

$$\frac{1}{r+\Lambda} \leq (\gamma-r)\phi \quad\Longleftrightarrow\quad 1 \leq (\gamma-r)(r+\Lambda)\phi \quad\Longleftrightarrow\quad \phi > \frac{1}{r+\Lambda(\gamma-r)}.$$

Thus, when $1 > (\gamma-r)(r+\Lambda)\phi$, the lender never sells its entire stake in finite time, i.e., $\lim_{t\to\infty}\beta_t = c^B > 0$ as well as $a^B > 0$.

A.6.2 Part 2

It remains to check whether V_t reaches 0 in finite time in which case W_t and c_t reach zero in finite time; this implies then under our implementation that the agent sells its entire stake in finite time, in that the implementation stipulates $\beta(0) = 0$.

Recall $\dot{V} = G(V)$ with $G(V) = (\gamma+\lambda(V))V - W$, so that $\lim_{V\to 0}\dot{V} = 0$. As $\lim_{V\to 0} F(V) = F^B(q)$, we have

$$F'(0) := \lim_{V\to 0} F'(V) = \frac{F^B(q) - (\gamma-r)\phi}{\phi} < 0,$$

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

Because of $a^B = a(V^B) = 0$, we have $a(V) < F(V)/\phi$ as well as

$$F''(V) = \frac{-(\gamma-r)F'(V)}{\dot{V}}. \tag{A.56}$$

in a right-neighbourhood of zero $(0, \hat{\epsilon})$ for appropriate $\hat{\epsilon} > 0$. Also recall that in this right-

neighbourhood

$$W(V) = a(V)\phi = F(V) - F'(V)[V + \phi] - (\gamma - r)\phi$$

and $a(V)$ are also differentiable.

We can calculate

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

where the second equality uses (A.56). As such, we can calculate

$$\lim_{V \to 0} W'(V) = +\infty$$

as well as

$$\lim_{V \to 0} \frac{\partial \dot{V}}{\partial V} = \lim_{V \to 0} \frac{\partial G(V)}{\partial V} = \lim_{V \to 0} G'(V) = \lim_{V \to 0} \Big(\gamma + \lambda(V) - W'(V)(1 + V/\phi)\Big) = -\infty.$$

It follows that $G(V)$ is not continuously differentiable on $[0, V_0]$ and thus is also not Lipschitz continuous in the same interval.

Next, notice that $G(V) < -V \iff G(V)/(-V) > 1$ on an interval $(0, \epsilon')$. This follows from the fact that

$$\lim_{V \to 0} \frac{G(V)}{-V} = \lim_{V \to 0} \frac{G'(V)}{-1} = \infty,$$

and continuity of $G(V)$ for $V > 0$, where we used L'Hopital's rule.

As a next step, we show that for $\alpha \in (0, 1)$ there exists $0 < \epsilon < \epsilon'$ such that on $(0, \epsilon)$:

$$G(V) < -V^\alpha \iff \frac{-V^\alpha}{G(V)} < 1.$$

To do so, we calculate

$$\begin{aligned} 0 \leq \lim_{V \to 0} \frac{-V^{-\alpha}}{G(V)} &= \lim_{V \to 0} \frac{-\alpha V^{\alpha-1}}{G'(V)} = \lim_{V \to 0} \frac{\alpha V^{\alpha-1}}{\frac{(\gamma-r)F'(V)}{G(V)}} = \lim_{V \to 0} \frac{-\alpha V^{\alpha-1} G(V)}{(\gamma - r)F'(V)} = \lim_{V \to 0} \frac{-\alpha V^{\alpha-1} G(V)}{(\gamma - r)F'(0)} \\ &= \lim_{V \to 0} \frac{\alpha V^{\alpha-1} G(V)}{(\gamma - r)F'(0)} \leq \lim_{V \to 0} \frac{\alpha V^{\alpha-1}(-V)}{(\gamma - r)F'(0)} = \lim_{V \to 0} \frac{\alpha V^\alpha}{-(\gamma - r)F'(0)} = 0, \end{aligned}$$

where we used L'Hopital's rule in the first equality and that $G(V) < -V$ in a neighbourhood of zero (and thus in the limit $V \to 0$) in the second last inequality. The remaining steps carry out simplifying calculations. Thus, by continuity, we have $G(V) < -V^\alpha < 0$ on $(0, \epsilon)$.

Let $T = \inf\{t \geq 0 : V_t = \epsilon\}$. As $\epsilon > 0$ and $G(V) < 0$ for $V \geq \epsilon$, it readily follows that T is finite, i.e., $T < \infty$.

Next, for times $t \geq T$, consider the ODE $X_t = -X_t^\alpha$ with $X_T = K > 0$ for $\alpha \in (0, 1)$ which

admits the general solution:[34]

$$X_t = \begin{cases} \left[K^{1-\alpha} - (1-\alpha)(t-T)\right]^{\frac{1}{1-\alpha}} & \text{for} \quad t < T' \\ 0 & \text{for} \quad t \geq T' \end{cases}$$

for a constant K. Thus, $X_T = K > 0$. It follows that X_t reaches 0 at time $T' = T + \frac{K^{1-\alpha}}{1-\alpha} < \infty$.

Set $K = \epsilon$, so that $X_T = V_T = \epsilon$. Due to $G(V) \leq -V^\alpha$ as well as $\dot{V}_t = G(V_t)$ and $\dot{X}_t = -X_t^\alpha$, it follows that $V_t \leq X_t$ for $t \geq T$. As such, V_t reaches 0 in finite time $T'' \leq T' < \infty$. Thus, in the implementation, β_t reaches $\beta(0) = \beta^B = 0$ in finite time, which was to show.

B Additional Results

B.1 Model variant with different default intensity

We now assume a different specification for the default rate λ_t, in that

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

and $\frac{\partial^2 \lambda_t}{\partial a_t q} = -\alpha$, where α captures whether screening and monitoring are substitutes ($\alpha < 0$) or complements ($\alpha > 0$) in reducing default risk. The baseline model is obtained upon setting $\alpha = 0$. A micro-foundation for this default rate can be found in Section B.6. As in the baseline, we first fix the level of q and solve the model for a given level of q, and finally determine optimal q.

Continuation value and monitoring incentives. Under this alternative specification of the default rate, the model solution and solution technique remain analogous to the ones of the baseline, but the formulae change in some instances. We sketch the solution and heuristically derive the relevant equations that characterize the optimal contract. To begin, recall that continuation value is given in (4) and follows (11), i.e., $\dot{W}_t = (\gamma + \lambda_t)W_t + \frac{\phi a_t^2}{2} - c_t$ for smooth payouts $dC_t = c_t dt$ after time zero. The agent chooses monitoring effort according to

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

so that provided $a_t \in (0, \bar{a})$:

$$a_t = \frac{(1+\alpha q)W_t}{\phi}.$$

The screening incentive condition can be written as $V_0 = \kappa q$, just as in the baseline.

[34] For a verification of this "guess," simply calculate for $t < T'$:

$$\dot{X}_t = -\left[K^{1-\alpha} - (1-\alpha)(t-T)\right]^{\frac{1}{1-\alpha}-1} = \left[K^{1-\alpha} - (1-\alpha)(t-T)\right]^{\frac{\alpha}{1-\alpha}} = -V_t^\alpha.$$

Screening incentives. Next, to derive the law of motion of V_t, we notice $\frac{\partial c_t}{\partial q} = 0$ (i.e., changes in hidden screening effort do not affect contracted payouts) and differentiate (11) with respect to q (using the envelope theorem, i.e., taking optimal monitoring a_t as given) to obtain:

$$\dot{V}_t = (\gamma + \lambda_t)V_t - (1 + \alpha a_t)W_t,$$

so that

$$V_t = \int_t^\infty e^{-\gamma(s-t)-\int_t^s \lambda_u du}(1 + \alpha a_s)W_s ds$$

after integrating over time. As before, the state space is $(V^B(q), V_0]$ with $\lim_{t\to\infty} V_t = V^B = V^B(q)$.

HJB equation and boundary conditions As in the baseline model, we conjecture and verify that $W_t = W$ is a control variable while $V_t = V$ is the only state variable in the dynamic optimization, in that total surplus is a function V_t only (i.e., $F_t = F(V_t)$). We omit time subscripts unless necessary. Invoking the dynamic programming principle and using the integral representation of total surplus in (20), we obtain that $F(V)$ solves on $(V^B(q), V_0]$ the HJB equation

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

which is very similar to the baseline HJB equation (23). We can use $a = \frac{W(1+\alpha q)}{\phi} \iff W = \frac{\phi a}{1+\alpha q}$ to eliminate W from the HJB equation, which yields

$$rF(V) = \max_a \left\{1 - \frac{\phi a^2}{2} - (\gamma - r)\left(\frac{\phi a}{1 + \alpha q}\right) - \lambda F(V) + F'(V)\left[(\gamma + \lambda)V - (1 + \alpha a)\left(\frac{\phi a}{1 + \alpha q}\right)\right]\right\},$$

subject to $W = \frac{\phi a}{1+\alpha q} \in [0, F(V)]$, i.e., the principal's limited liability constraintc, and $a \in [0, \bar{a}]$.

The first order condition with respect to a reads

$$(1 + \alpha q)(F(V) - F'(V)V) - \phi a - \frac{\phi(\gamma - r)}{1 + \alpha q} - F'(V)\left(\frac{\phi}{1 + \alpha q} + \frac{2\alpha a\phi}{1 + \alpha q}\right) = 0.$$

Thus, if optimal effort is interior, we have:

$$a = a(V) = \frac{(1 + \alpha q)(F(V) - F'(V)V) - \frac{\phi(\gamma-r)}{1+\alpha q} - \frac{F'(V)\phi}{1+\alpha q}}{\phi + \frac{2\alpha\phi F'(V)}{1+\alpha q}}.$$

Finally, the boundary condition for above HJB arises arises when $V_t \to V^B$ and $\dot{V}_t \to 0$, where

$$V^B = V^B(q) = \frac{(1 + \alpha a^B)W^B}{\gamma + \lambda^B}$$

and $\lambda^B = \Lambda - a^B - q - \alpha a^B q$ as well as $a^B = \frac{1+\alpha q)W^B}{\phi}$. The optimal level of a^B is determined

according to:

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

so that

$$a^B = \max\left\{ 0, \frac{(1+\alpha q)F^B - \frac{\phi(\gamma - r)}{1+\alpha q}}{\phi} \right\}$$

Note that a^B may be equal to zero, e.g., when ϕ is sufficiently large.

Optimal screening The screening incentive condition is $V_0 = \kappa q$. The total surplus function, given q, is characterized above and denoted $F(V)$. Optimal q is determined to maximize

$$\max_{q \in [0,\bar{q}]} F(V_0) \quad \text{s.t.} \quad V_0 = \kappa q,$$

which is entirely analogous to the baseline. As in the baseline, we generally have $V_0 > V^B$ and V_t drifts down over time, i.e., $\dot{V}_t < 0$ with $\lim_{t\to\infty} \dot{V}_t = 0$.

B.2 Loan Portfolio

We now allow the lender to originate a portfolio of loans which consists, for simplicity, of two ex-ante identical loans indexed by $i \in \{1, 2\}$. Unless otherwise mentioned, the assumptions underlying this model variant as well as players' preferences remain as in the baseline model. Each loan i pays coupons at rate 1 up to its time of default τ^i. Each loan i defaults with the time-varying intensity

$$\lambda_t^i = \Lambda - q^i - a_t^i,$$

where q^i is the lender's screening of loan i at time $t = 0^-$ and a_t^i is the lender's monitoring of loan i at time t. The two loans' random default times are independent, conditional on the lender-investor contract $\mathcal{C}$. Both efforts are bounded, i.e., $a_t^i \in [0, \bar{a}]$ and $q^i \in [0, \bar{q}]$.

The lender-investor contract $\mathcal{C} = (\hat{q}^i, \hat{a}_t^i, dC_t)$, signed at time $t = 0^-$, stipulates incremental payouts dC_t as well as recommended screening and monitoring $(\hat{q}^i, \hat{a}_t^i)$ for both loans i. The costs of screening and monitoring are, as in the baseline, quadratic and separable across loans. Given $\mathcal{C}$, the lender's payoff at time $t = 0^-$ (i.e., before screening) reads

$$W_{0^-} = \max_{q, \{a_t\}} \mathbb{E}\left[\int_0^\infty e^{-\gamma t} \left(dC_t - \frac{\phi(a_t^1)^2}{2} dt - \frac{\phi(a_t^2)^2}{2} dt \right) \right] - \frac{\kappa(q^1)^2}{2} - \frac{\kappa(q^2)^2}{2}, \tag{B.57}$$

with exogenous screening and monitoring cost parameters $\kappa, \phi \geq 0$. After screening at time $t = 0^-$, the lender's continuation payoff for $t \geq 0$ reads

$$W_t := \mathbb{E}\left[\int_t^\infty e^{-\gamma(s-t)} \left(dC_s - \frac{\phi(a_s^1)^2}{2} ds - \frac{\phi(a_s^2)^2}{2} ds \right) \right]. \tag{B.58}$$

As in the baseline model, we focus on incentive-compatible contracts, so that $a_t^i = \hat{a}_t^i$ and $q^i = \hat{q}^i$.

Monitoring effort a_t^i becomes zero after loan i defaults. As both loans are identical, we also focus on contracts that implement symmetric screening and monitoring efforts, i.e., $\hat{q}^1 = \hat{q}^2$ and $\hat{a}_t^1 = \hat{a}_t^2$ (before the first time of default $\min\{\tau^1, \tau^2\}$).

The principal's payoff reads

$$F_t = \mathbb{E}_t \left[\int_t^\infty e^{-r(s-t)} \left(X_s ds - dC_s \right) \right],$$

where

$$X_s = \begin{cases} 2 & \text{for} \quad s < \min\{\tau^1, \tau^2\} \\ 1 & \text{for} \quad s \in \left[\min\{\tau^1, \tau^2\}, \max\{\tau^1, \tau^2\} \right] \\ 0 & \text{for} \quad s > \max\{\tau^1, \tau^2\} \end{cases}$$

is the loan portfolio's cash flow. Analogous to (20), total continuation surplus can then be written as

$$F_t = \mathbb{E}_t \left[\int_t^\infty e^{-r(s-t)} \left(X_s ds - \frac{\phi(a_s^1)^2}{2} ds - \frac{\phi(a_s^2)^2}{2} ds - (\gamma - r) W_s ds \right) \right]. \tag{B.59}$$

The optimal contract chosen at time $t = 0^-$ maximizes total continuation surplus at time $t = 0^-$

$$F_0^- = F_0 - \frac{\kappa(q^1)^2}{2} - \frac{\kappa(q^2)^2}{2},$$

subject to all relevant incentive constraints (derived shortly) and the lender's and investors' limited liability constraints. As in the baseline model, we assume for convenience that parameters are such that optimal efforts are interior.

In what follows, we provide the heuristic solution of this model variant, assuming that all necessary regularity conditions (e.g., for solution existence) are met and the first order approach is valid. Before proceeding, note that it is always possible to incentivize both loans separately, for instance, by signing the baseline contract which we denote for convenience by $\mathcal{C} = (q^{Base}, a_t^{Base}, dC_t^{Base} = c_t^{Base})$ for both contracts. In our notation of this variant, this would be achieved via $\hat{q}^i = q^{Base}$, $\hat{a}_t^i = a_t^{Base}$ (where effort becomes zero once the respective loan defaults), and $dC_t = 2c_t^{Base} dt$. The notable result of the following analysis is that structuring lender compensation on the portoflio level generally improves upon the baseline contract and thus does better in terms of incentives. The intuition is that structuring the lender's compensation on the portfolio level relaxes loan-level limited liability, hence facilitating more efficient incentive provision.

B.2.1 Optimal Contract for Loan Portfolio

We start by solving the model, taking $q = q^1 = q^2$ as given. We denote now by $\tau' = \min\{\tau^1, \tau^2\}$ the time of the first default and by $\tau = \max\{\tau^1, \tau^2\}$ the time of the second default. It is natural to conjecture that, upon any default, the agent looses its entire continuation payoff/stake, so as to provide incentives to screen and monitor effectively. Specifically,

$$\lim_{t \uparrow \hat{\tau}} W_t \geq W_{\hat{\tau}} = 0 \quad \text{for} \quad \hat{\tau} \in \{\tau', \tau\}.$$

Solution after time τ'. Let us consider $t > \tau'$ and conjecture that $dC_t = c_t dt$. Suppose that loan i has not defaulted yet while loan $-i$ has defaulted, where we adopt the notation that $i = 1$ ($i = 2$) implies $-i = 2$ ($-i = 1$). The model solution after time τ' is akin to our baseline solution (modulo starting values). We obtain

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

As in the baseline, the lender chooses a_t^i to maximize $-\lambda_t^i W_t - \frac{\phi(a_t^i)^2}{2}$, so that $a_t^i = \frac{W_t}{\phi}$ for $t > \tau'$.

Next, let us define $V_t^{Post} = \frac{\partial W_t}{\partial q^i}$ for $t > \tau'$. As in the baseline, we have for $t > \tau'$:

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

or

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

As in the baseline model, we conjecture the optimal continuation contract after time τ' is fully characterized by the state variable $V^{Post} = V_t^{Post}$. Thus, total surplus F_t, continuation payoff W_t, monitoring a_t^i, and payouts c_t can be written as functions of $V^{Post} = V_t^{Post}$ only, in that $F_t = F^{Post}(V^{Post})$, $W_t = W^{Post}(V^{Post})$, $a_t^i = a^{Post}(V^{Post})$, and $c_t = c^{Post}(V^{Post})$. We omit time subscripts unless necessary.

That is, V^{Post} is the state variable in the dynamic optimization, while $W = W^{Post}$ becomes a control variable. Using (B.59), total surplus $F^{Post}(V^{Post})$ satisfies the HJB equation:

$$rF^{Post}(V^{Post}) = \max_{a^{Post}\in[0,\bar{a}], W^{Post}} \Big\{ 1 - \frac{\phi(a^{Post})^2}{2} - (\gamma - r)W^{Post} - \lambda^{Post} F^{Post}(V^{Post}) + (F^{Post})'(V^{Post})\big((\gamma + \lambda^{Post})V^{Post} - W^{Post}\big) \Big\} \quad \text{(B.60)}$$

subject to $a^{Post} = W^{Post}/\phi$ and $W^{Post} \in [0, F^{Post}(V^{Post})]$ and with $\lambda^{Post} = \Lambda - a^{Post} - W^{Post}$.

To characterize the boundary behavior, let

$$V^{B,Post} = \frac{W^{Post}(V^{B,Post})}{\gamma + \lambda^{B,Post}}$$

with $\lambda^{B,Post} = \Lambda - a^{Post}(V^{B,Post}) - q$. HJB equation (B.60) is solved for $V^{Post} > V^{B,Post}$ subject to the boundary condition

$$\lim_{V^{Post}\to V^{B,Post}} F^{Post}(V) = F^{B,Post},$$

with

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

subject to $a^{Post} = W^{Post}/\phi$.

Lastly, having solved for optimal effort $a^{Post}(V^{Post})$ and continuation payoff $W^{Post}(V^{Post})$, we

obtain for $t > \tau'$

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

that is,

$$c^{Post}(V^{Post}) = (\gamma + \lambda^{Post}(V^{Post}))W^{Post}(V^{Post}) - (W^{Post})'(V^{Post})\dot{V}^{Post} + \frac{\phi(a^{Post}(V^{Post}))^2}{2}.$$

Solution before time τ'. Consider times $t < \tau'$, so neither loan has defaulted yet. Recall that upon the first default event at time τ', the agent looses its entire stake. Further, conjecture and verify that payouts are smooth prior to time τ', i.e., $dC_t = c_t dt$ for $t < \tau'$. At time τ', there can be a (negative) lump-sum payment $dC_t^{Post} := dC_{\tau'}$ to the agent, to be clarified later. Then, we obtain

$$\dot{W}_t = (\gamma + \lambda_t^1 + \lambda_t^2)W_t - c_t + \frac{\phi(a_t^2)^2}{2} + \frac{\phi(a_t^2)^2}{2} - (\lambda_t^1 + \lambda_t^2)(dC_t^{Post} + W_t^{Post}), \tag{B.61}$$

where it is optimal to set $dC_t^{Post} + W_t^{Post} = 0$. That is, upon first default at time $t = \tau'$, the lender looses its entire stake. To gain a potentially positive new stake $\lim_{s\downarrow\tau'} W_s \geq 0$, the agent makes a (possibly negative) lump-sum payment of dC_t^{Post} to the investors.

The lender chooses a_t^i to maximize

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

so that $a_t^i = \frac{W_t}{\phi}$. As such, $\lambda_t^i = \Lambda - q^i - a_t^i = \Lambda - q^i - \frac{W_t}{\phi}$.

Next, we define $V_t^i = \frac{\partial}{\partial q^i} W_t$ for $t < \tau$ and $i = 1, 2$. We assume that the first order approach is valid. As such, using the lender's objective at time $t = 0^-$ in (B.57), the screening incentive conditions become

$$V_0^i = \kappa q^i \quad \text{for} \quad i = 1, 2.$$

We now characterize the dynamics of V_t^i (for $t < \tau$). To do so, note that $\frac{\partial dC_t^{Post}}{\partial q^i} = 0$ as the contracted lump-sum payment at the first time of default cannot depend on hidden screening. Likewise, $\frac{\partial c_t}{\partial q^i} = 0$. Second, observe that $\frac{\partial W_t^{Post}}{\partial q^i} = V_t^{Post} \mathbb{I}\{\tau^i > \tau'\}$, i.e., the impact of screening effort q^i lasts beyond time τ' if and only if loan i is not the loan that defaults first. Here, $\mathbb{I}\{\cdot\}$ is the indicator function which equals one if $\{\cdot\}$ is true and zero otherwise. With these insights in mind, we can differentiate the law of motion of W_t from (B.61) with respect to q^i (using the envelope theorem, i.e., taking optimal monitoring a_t^i as given). After some algebra, we obtain

$$\dot{V}_t^i = (\gamma + \lambda_t^1 + \lambda_t^2)V_t^i - W_t - \lambda_t^{-i} V_t^{Post}.$$

That is,

$$V_t^i = \int_t^s e^{-\gamma(s-t) - \int_t^s (\lambda_u^1 + \lambda_u^2)du} (W_s + \lambda_s^{-i} V_s^{Post}) ds. \tag{B.62}$$

Because the two loans are identical and we focus on symmetric screening and monitoring of these two loans prior to time τ', we have $V_t^1 = V_t^2 \equiv V_t$ for $t < \tau'$ and $q = q^1 = q^2$.

Using $a_t^i = W_t/\phi$ as well as each loan's default intensity $\lambda_t \equiv \lambda_t^i = \Lambda - q^i - a_t^i = \Lambda - q^i - \frac{W_t}{\phi}$,

we obtain

$$\dot{V}_t = (\gamma + 2\lambda_t)V_t - W_t - \lambda_t V_t^{Post}.$$

As in our previous analysis, we conjecture and verify that prior to time τ', all payoff-relevant quantities can be written as functions of $V = V_t$ only, in that $F_t = F(V)$, $W_t = W(V)$, $a_t^i = a(V)$, and $c_t = c(V)$. We omit time subscripts, unless necessary.

Using (A.27) as well as invoking the dynamic programming principle, $F(V)$ satisfies:

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

subject to $W \in [0, F(V)]$ and $a = W/\phi$. Using $W = a\phi$, we can rewrite this HJB equation to get

$$rF(V) = \max_{a,V^{Post}} \Bigg\{ 2 - \phi a^2 - (\gamma - r)\phi a + 2\lambda(F^{Post}(V^{Post}) - F(V)) \\ + F'(V)\Big((\gamma + 2\lambda)V - \phi a - \lambda V^{Post}\Big) \Bigg\}$$

Note that V^{Post} —which is determined b by the continuation contract after time τ' — is a choice variable that affects screening incentives before time τ' and at time $t = 0$. The derivative with respect to V^{Post} reads

$$2(F^{Post})'(V^{Post}) - F'(V).$$

Provided there exists an interior solution, we then have

$$2(F^{Post})'(V^{Post}) - F'(V) = 0,$$

which pins down V^{Post} as a function of V, i.e., $V_t^{Post} = V^{Post}(V)$. We assume this is the case; we verify this outcome in our numerical solution procedure.

Next, provided it is interior, optimal effort a satisfies the first order condition

$$0 = -2\phi a - (\gamma - r)\phi + 2(F(V) - F^{Post}(V^{Post})) - 2F'(V)V - F'(V)\phi + F'(V)V^{Post}$$

so that

$$a(V) = \min\left\{ \frac{2(F(V) - F^{Post}(V^{Post})) - (\gamma - r)\phi - 2F'(V)V - F'(V)\phi + F'(V)V^{Post}}{2\phi}, \frac{F(V)}{\phi} \right\}.$$

Thus, $W(V) = \phi a(V)$.

Finally, we characterize the boundary behavior of the HJB equation. For this sake, consider

$$V^B = \frac{W(V^B) + \lambda^B V^{Post}}{\gamma + 2\lambda^B},$$

with $\lambda^B = \Lambda - q - a(V^B)$. Above HJB equation is solved for $V > V^B$ subject to the boundary condition

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

with

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

subject to $a = W/\phi$.

Having solved the optimal contract for $t < \tau'$, we obtain payouts $c_t = c(V_t)$

$$c(V) = (\gamma + 2\lambda(V))W(V) + \phi a(V)^2 - W'(V)\dot{V}.$$

Optimal screening effort at time $t = 0^-$ is then determined according to:

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

We note that the continuation surplus at $t = 0$, i.e., $F_0 = F(V_0)$, depends on $q = q^i$.

B.2.2 Implementation

We now discuss the implementation of the optimal contract. For this sake, consider times $t < \tau'$ and $t > \tau'$ separately. After the first default, the implementation becomes analogous to the one from the baseline. Specifically, for $t > \tau'$, we define

$$L_t^{Post} = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_s^{Post}} 1 ds$$

as the market value of the loan. Then, the lender's retention β_t^{Post} is determined according to

$$c_t^{Post} = c^{Post}(V_t^{Post}) = \beta_t^{Post} - \dot{\beta}_t^{Post} L_t^{Post}.$$

The boundary condition is $\lim_{t\to\infty} \dot{\beta}_t^{Post} = 0$. We assume that a unique (non-constant) solution to this ODE exists. As in the baseline model, we expect $\dot{\beta}_t^{Post} < 0$.

Next, consider t before the first time of default, $t < \tau'$. We now offer an implementation in which the loan portfolio is tranched into an equity and safe tranche. The lender only retains the equity tranche. Suppose that the equity tranche is in net supply $E > 0$. The equity tranche (in supply E) pays cash flows of one only up to the first time of default τ'; the safe tranche is implicitly defined as the residual tranche and discussed in more detail below. Specifically, one unit of equity tranche pays cash flows at rate 1 up to τ', leading to market value/price:

$$L_t^E = \int_t^\infty e^{-r(s-t)-2\int_t^s \lambda_u du}.$$

The contracted payouts to the agent prior to time τ' read

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

Then, we get the retention level of the equity tranche via:

$$\beta_t^E + (-\dot{\beta}_t^E)L_t^E = c_t.$$

The boundary condition is $\lim_{t\to\infty} \dot{\beta}_t^E = 0$. We assume that a unique (non-constant) solution to this ODE exists. Again, notice that the contracted payments c_t are implemented by having the agent retain a time-varying share of the equity tranche β_t^E.

We still need to determine E, the supply of the equity tranche. It is natural to set $E = \max\{1, \beta_0^E\}$, so that the total supply of the equity tranche E is bigger than the lender's holding at time $t = 0$. Next, we normalize the supply of the safe tranche to 1. The safe tranche then pays cash flows $2 - E$ per unit before time τ' and cash flows of 1 per unit for times $t \in [\tau', \tau]$. The safe tranche becomes worthless at time τ (i.e. when both loans have defaulted).

At time $t = \tau'$, the equity tranche is wiped out (i.e., its value drops to zero), while 1 unit of the safe tranche remains. The market value of the safe tranche equals L_t^{Post} for $t \in (\tau', \tau)$ and

$$L_t^S = \int_t^\infty e^{-r(s-t)-2\int_t^s \lambda_u du} \Big(2 - E + 2\lambda_s L_s^{Post}\Big) ds$$

for time $t < \tau'$. Thus, in general, the equity tranche changes its value at time τ'.

At time $t = \tau'$, the lender needs to buy $\beta_{\tau'}^{Post}$ units of the loan to rebuild its stake. The lender buys at price $W_{\tau'}^{Post}/\beta_{\tau'}^{Post}$ so that it pays in total $-dC_{\tau'}^{Post} = W_{\tau'}^{Post}$ at time τ' in exchange for continuation value $W_{\tau'}^{Post}$.

B.2.3 Analysis

One way to incentivize the origination of the two identical loans is to contract for each loan separately by utilizing the baseline contract (which is optimal on the individual loan level). The interpretation of this arrangement is that the lender retains a share of each individual loan and sells off these shares over time. Crucially, if loan i defaults, the agent's stake in loan i is wiped out, but the value of the agent's stake in loan $-i$ is (not directly) affected and maintains value. That is, the agent is de facto protected by limited liability on the loan level.

In contrast, the optimal contract for loan portfolios stipulates that if loan i defaults, the agent is punished by loosing its entire stake, so that the contract does not respect loan-level limited liability but merely portfolio-level limited liability. By construction, the optimal contract for loan portfolios does at least weakly better than contracting separately for each loan. We now numerically compare the outcomes of the optimal contract for loan portfolios and the baseline contract (i.e., separate contracts for each loan). To this end Figure B.1 plots initial monitoring and screening efforts against the parameters ϕ, κ, Λ, and γ both under the baseline (solid black line) and when lender compensation is structured on the portfolio level. It can be seen that, as expected, the optimal contract for loan portfolios incentivizes higher screening and monitoring efforts across than the baseline contract, notably, across all parameters considered. A corollary of this observation is that, because the two contracts differ in that way, the optimal contract for loan portfolios leads to higher surplus. The intuition is that structuring lender compensation on the portfolio rather than the individual loan level relaxes loan-level limited liability and thus facilitates the more efficient provision of screening and monitoring incentives. An applied insight of our analysis is that optimal incentive provision for a portfolio of loans involves tranching in a way that the lender mostly retains the riskier tranche.

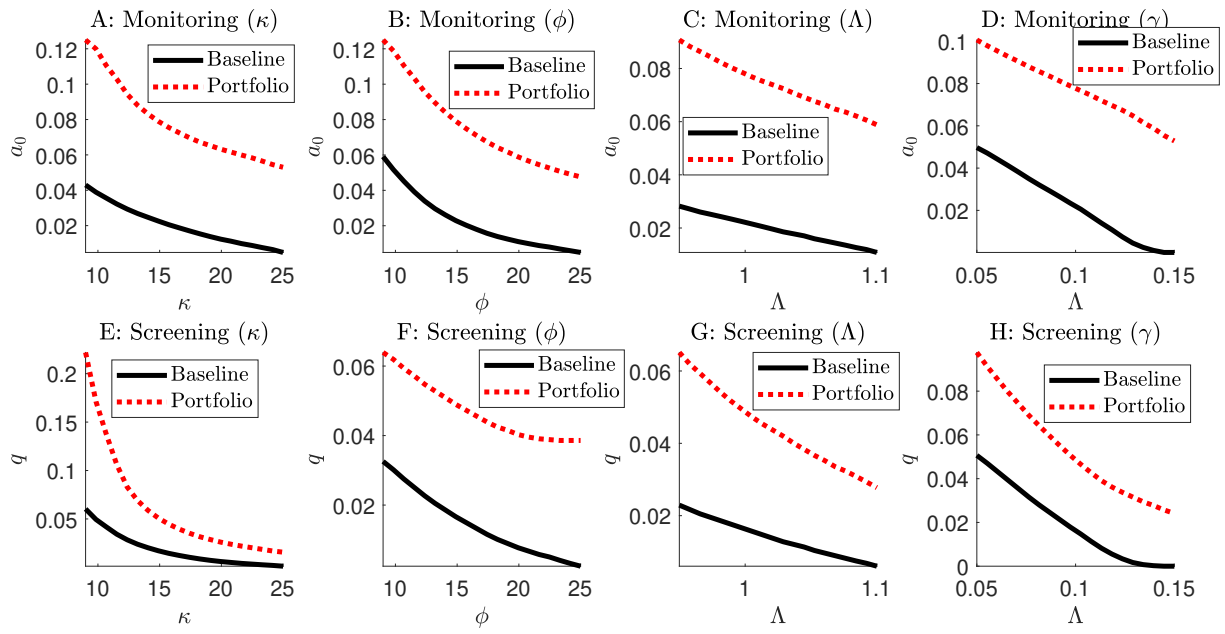

Figure B.1: **Comparative Statics and Loan Portfolios.** This figure plots monitoring effort a_0 and screening effort q against the parameters ϕ, κ, Λ, and γ both under the baseline (solid black line) and when lender compensation is structured on the portfolio level (dotted red line). We use our baseline parameters, but consider higher cost of screening and monitoring (i.e., $\phi = \kappa = 15$) to ensure optimal efforts are interior also when considering portfolios.

B.3 Model Extension with Finite Maturity

B.3.1 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, 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 proability 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:

$$\dot{W}_t = (\gamma + \delta + \lambda) W_t + \frac{\phi a_t^2}{2} - c_t - \delta dC_t^\delta. \tag{B.63}$$

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

$$(\gamma + \delta) W_t = \max_{a_t \in [0, \bar{a}]} \left(c_t - \lambda_t W_t - \frac{\phi a_t^2}{2} + \delta dC_t^\delta + \dot{W}_t \right), \tag{B.64}$$

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 (B.64) with respect to q to obtain for $V_t = \frac{\partial}{\partial q} W_t$:[35]

$$\dot{V}_t = (\gamma + \delta + \lambda_t) V_t - W_t. \tag{B.65}$$

[35] An alternative derivation (not relying explicitly on envelope theorem) simply rewrites (B.63) 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

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

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 by

$$F_t = \int_t^\infty e^{-(r+\delta)(s-t)-\int_t^s \lambda_u du} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s + \delta F_s^\delta\right) ds. \tag{B.66}$$

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.

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

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

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

$$(r+\delta)F(V) = \max_{a,W} \left\{1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + \delta F^\delta + F'(V)\big((\gamma+\delta+\lambda)V - W\big)\right\}, \tag{B.67}$$

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

As in the baseline, the state variable V_t converges to a limit $V^B(q)$, i.e., $\lim_{t\to\infty} V_t = V^B(q)$, whereby $\lim_{t\to\infty} \dot{V}_t = 0$.[36] Then, the HJB equation (B.67) is subject to the boundary condition

$$\begin{aligned}\lim_{V \to V^B(q)} F(V) = F^B(q) = & \\ & \max_{W \in [0, F^B(q)]} \left(\frac{1 + \delta F^\delta}{r + \Lambda - a - q + \delta} - \frac{(\gamma - r)W}{r + \Lambda - a - q + \delta} - \frac{\frac{\phi a^2}{2}}{r + \Lambda - a - q + \delta}\right),\end{aligned} \tag{B.68}$$

[36] We 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.

which is analogous to (24) in the baseline model. Here,

$$V^B(q) = \frac{W^B(q)}{r+\delta+\Lambda-a^B-q} \quad \text{with} \quad W^B(q) = W(V^B(q)) \quad \text{and} \quad a^B(q) = \frac{W^B(q)}{\phi}. \tag{B.69}$$

We assume that a unique solution to (B.67) (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 (B.66) 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.[37] 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.[38]

Thus, using $F^\delta = F(V)$, the HJB equation (B.67) simplifies to

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

with $W = \phi a$ and $W \leq F$ (limited liability). The boundary condition (B.68) simplifies to

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

Optimal effort becomes

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

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.

B.3.2 Implementation

Payouts to the agent read by (B.63)

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

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

[38] In reality, loans mature deterministically and this feature naturally holds, preventing arbitrage.

Finally, we define the retention level β_t via

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

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, $c_t = c(V_t)$. Or

$$c(V) = (\gamma + \delta + \lambda(V))W(V) + \frac{\phi a(V)^2}{2} - W'(V)\dot{V} - \delta dC^\delta.$$

Set $dC^\delta = \beta(V)F(V)$ (i.e., $dC_t^\delta = \beta_t F_t$) so that upon maturity, the agent receives fraction $\beta(V)$ of the payout $F(V)$. 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 $F_s^\delta = F_s$ — that $L_s^\delta = L_s$, leading to $L_t = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}1ds$. That is, the maturity event is value neutral for total surplus $F(V)$ and the value of debt.

B.3.3 Main Results and Figures with Finite Maturity

We now replicate Figures 2 and 4 for finite maturity, where we choose $\delta = 0.2$, i.e., a maturity of 5 years which is close to the average maturity reported in Blickle et al. (2022) (that is, 4.43 years). Similar to Figure 2 in the baseline (infinite maturity) case, Figure B.2 plots screening and monitoring effort against κ, ϕ, Λ, and γ for different levels of α. Recall we use $\lambda_t = \Lambda - a_t - q - \alpha a_t q$. Indeed, as Figure B.2 illustrates, monitoring and screening efforts decrease with κ, ϕ, Λ, and γ for any α considered, producing qualitatively similar patters as Figure 2 does.

Next, similar to Figure 4 in the baseline (infinite maturity) case, Figure B.3 plots retention levels and selloff speed against κ, ϕ, Λ, and γ. Again, it can be seen that Figure B.3 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.

B.4 Repeated Interactions

Repeated lender-borrower and lender-investor interactions are common in credit markets, in particular in syndicated lending. To begin, note that our baseline setting already captures repeated lender-borrower interactions as it can be interpreted as follows: The lender extends a loan (with face value K and coupon payments at rate 1) to the borrower with a possibly finite maturity and this loan is rolled over (at identical terms and without re-screening) at maturity until default occurs. The cash flows from these repeated lender-borrower interactions are 1 up to default at time τ as the loan is simply rolled over at maturity dates.

Next, we analyze (possibly infinitely many) repeated lender-investor interactions with repeated loan origination. A key result of the analysis below is that repeated interactions facilitate lender commitment to a specific retention path stipulated in the contract implementation.

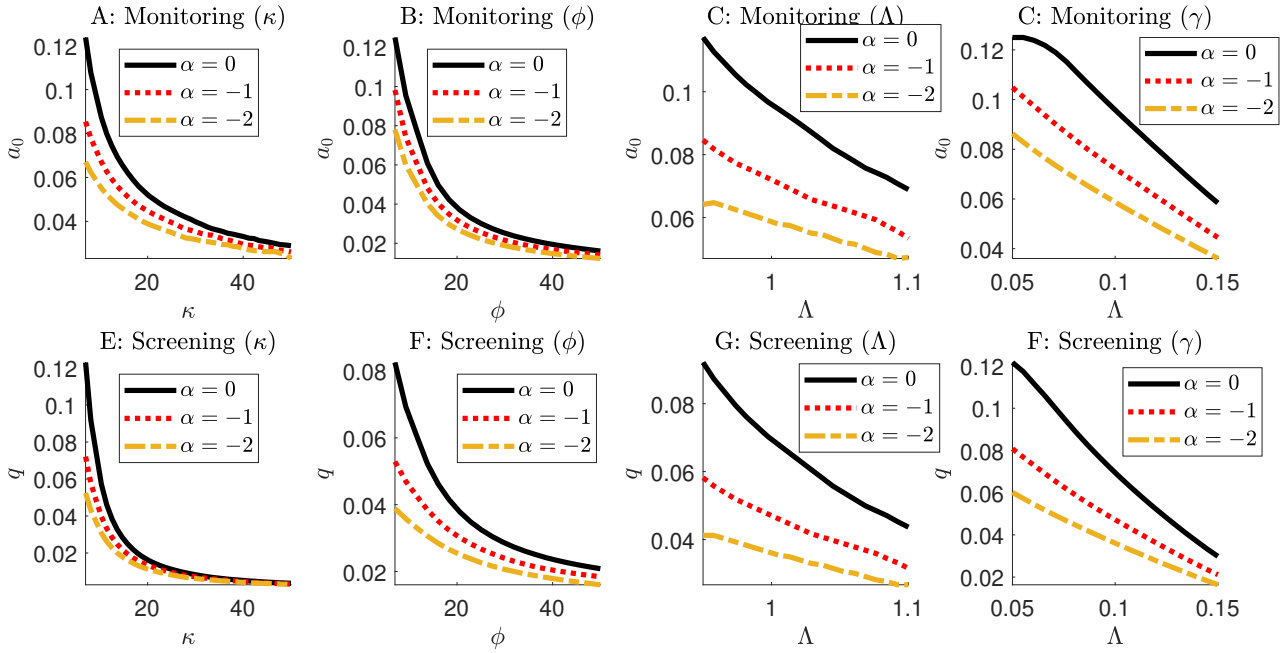

Figure B.2: **Comparative statics with finite maturity.** This figure plots monitoring effort a_t at $t = 0$ (solid black line), at $t = 5$ (dotted red line), and $t \to \infty$ (dashed yellow line) and screening effort q^* against the parameters ϕ, κ, Λ, and γ. We use our baseline parameters and set $\delta = 0.2$.

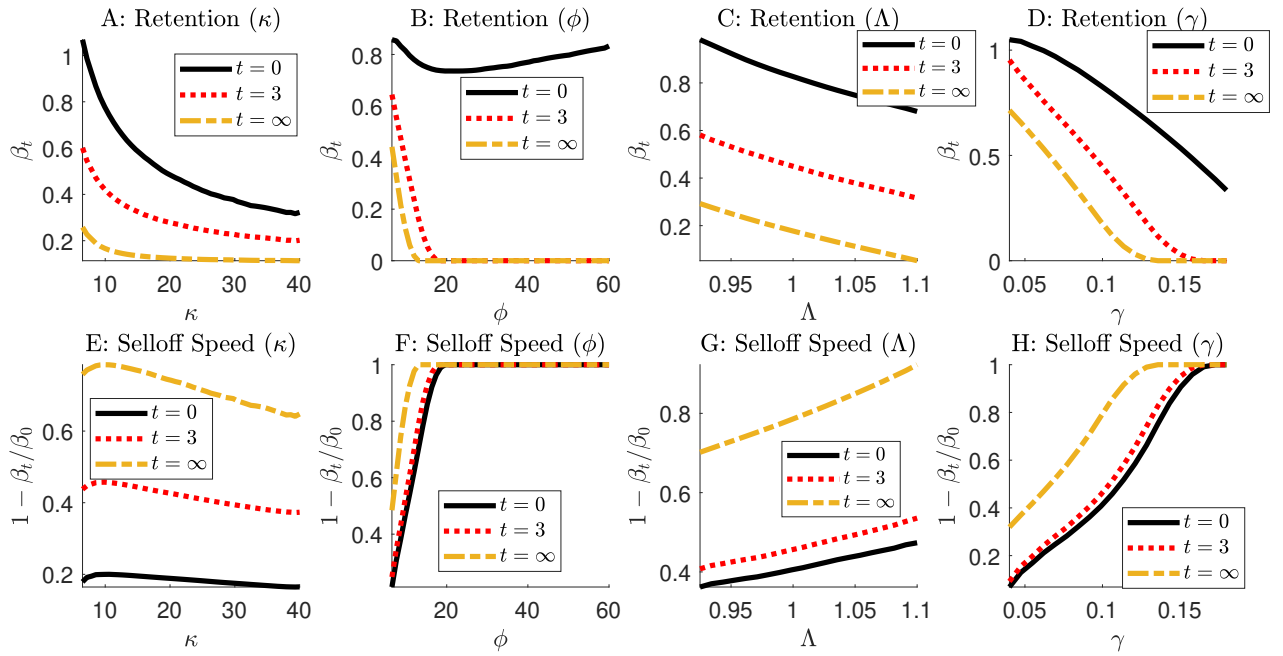

Figure B.3: **Retention and dynamics with finite maturity**. We use our baseline parameters and $\delta = 0.2$.

To formally analyze repeated lender-investor interactions, we consider that (after screening) the lender originates a loan with face value K, coupon payments at rate 1, and stochastic maturity arriving with exogenous intensity $\delta > 0$ at time time τ^δ. Here, we assume that the face value is not prohibitively large, so originating a loan has a positive net present value and is optimal for the lender. When the loan matures at time τ^δ, the face value is repaid, and the lender originates a new and identical loan to a new borrower with exogenous probability p^δ, so that the lender has to re-screen and exert costly screening effort when a new loan is originated. Otherwise, with probability $1 - p^\delta$, the lender exits and cannot originate further loans. If a loan defaults at time τ, the lender originates another loan with (exogenous) probability p^λ. With probability $1 - p^\lambda$, the lender exits (i.e., the relationship breaks down). We introduce the probabilities p^δ and p^λ to make the continuation of the relationship probabilistic, thereby capturing the fact that the number of repeated interactions is not infinite in practice. For simplicity, we assume that $p^\delta = 1$ and $p^\lambda \leq 1$. That is, the lender faces the possibility of being excluded from the secondary market upon default and not being able to originate further loans (due, e.g., to a loss in competitiveness related to the inability to sell the loan). Setting $p^\lambda = 0$ and $\delta = 0$ yields our baseline model.

We further assume that whenever the lender originates a loan, it receives an exogenous lump-sum reward $R \geq 0$, capturing the fees earned during the origination process. The parameter R is not central to the following arguments and can be set to zero without qualitatively changing the results. Nevertheless, it is useful to introduce R to illustrate that origination fees substitute for retention in incentive provision; thus, when retention is low in practice, the lender's incentives need not be low, as higher R can substitute for lower retention in incentive provision.

Unless otherwise mentioned, we maintain the assumptions of the baseline model, i.e., each loan's default intensity at time t is $\lambda_t = \Lambda - a_t - q$ where q is the screening effort and a_t is monitoring effort. At time $t = 0^-$, the lender and investors sign a long-term and full-commitment contract $\mathcal{C}$ stipulating payouts to the agent dC_t as well as recommended screening $\hat{q}$ and monitoring levels $\hat{a}_t$. We focus on incentive compatible contracts.

As in the previous proofs, we take the optimal level of q as given and characterize the continuation contract after screening q is chosen. This continuation contract yields continuation surplus F_t at time t. The initial level of screening is chosen to maximize $F_{0^-} = \max_{q \in [0,\bar{q}]} F_0 - \frac{\kappa q^2}{2}$, which is the payoff from loan origination excluding the lump-sum reward R and the face value K.

B.4.1 Solution and Optimal Contract

We now provide the heuristic solution for the lender-investor long-term and full-commitment contract, signed between investors and the lender at time $t = 0^-$ (i.e., before screening of the initial loan). In doing so, we take the screening level q as given and first determine the optimal continuation contract after screening. We also assume that relevant regularity conditions (so that a solution exists and is unique) are met and the first order approach is valid.

We conjecture and verify that outside loan origination, default, and maturity events, payouts to the agent are smooth at rate c_t. To begin, we define the agent's continuation value at time t as

$$W_t = \int_t^\infty e^{-(\gamma+\delta)(s-t)-\int_t^s \lambda_u du} \left(c_s - \frac{\phi a_s^2}{2} + \delta(W_s^\delta + dC_s^\delta) + \lambda_s p^\lambda (W_s^\lambda + dC_s^\lambda) \right) ds, \tag{B.70}$$

where W_s^δ (W_s^λ) is the agent's continuation payoff just after the loan matures (defaults and a new loan is originated) at time s and dC_s^δ (dC_s^λ) is a lump-sum payment that the agent receives when the loan matures (defaults) at time s and the lender originates a new loan. Taken together, the agent's total value jumps from W_s to $W_s^\delta + dC_s^\delta$ ($W_s^\lambda + dC_s^\lambda$) at maturity (default and new origination) at time s. Note that when the loan defaults and the lender cannot originate another loan (with probability $1 - p^\lambda$), its continuation payoff is zero and there is no lump-sum payment. Limited liability requires $W_s^\delta + dC_s^\delta \geq 0$ and $W_s^\lambda + dC_s^\lambda \geq 0$, i.e., the agent's continuation pay (including lump-sum transfer) is positive at maturity and default events. Limited liability (for the lender) also requires, as in the baseline model, $W_s \geq 0$ as well as $W_s^\delta, W_s^\lambda \geq 0$.

Next, we can differentiate (B.70) with respect to time t to obtain

$$\dot{W}_t = (\gamma + \delta + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \delta(W_t^\delta + dC_t^\delta) - \lambda_t p^\lambda (W_t^\lambda + dC_t^\lambda) - c_t. \tag{B.71}$$

The lender chooses monitoring effort to maximize

$$\max_{a_t \in [0,\bar{a}]} \left(\lambda_t \left[p^\lambda (W_t^\lambda + dC_t^\lambda) - W_t \right] - \frac{\phi a_t^2}{2} \right),$$

so that, provided monitoring effort is interior, we have

$$a_t = \frac{W_t - p^\lambda (W_t^\lambda + dC_t^\lambda)}{\phi}.$$

It is natural to conjecture that, to provide efficient and optimal incentives to screen and monitor, the optimal contract sets $(W_t^\lambda + dC_t^\lambda) = 0$, so as not to "reward" the agent for default. Thus, in what follows, we consider that $(W_t^\lambda + dC_t^\lambda) = 0$ so that $a_t = W_t/\phi$.

As a next step, we differentiate (B.71) with respect to q noting that $\frac{\partial dC_t^\delta}{\partial q} = 0$ (i.e., hidden screening does not affect contracted payouts) and $\frac{\partial W_t^\delta}{\partial q} = 0$ (i.e., the effects of screening effort impact do not extend beyond maturity) to obtain

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

so that $V_t = \int_t^\infty e^{-(\gamma+\delta)(s-t) - \int_t^s \lambda_u du} W_s ds$. The screening incentive condition is analogous to that in the baseline model, i.e., $V_0 = \kappa q$.

Next, we denote the continuation surplus after maturity (default and new origination) at a time s by F_s^δ (F_s^λ); if the loan defaults and origination is not possible, the continuation surplus becomes zero. As all originated loans are ex-ante identical and thus feature in optimum identical screening, we obtain $F_s^\delta = F_{0^-} + R$ and $F_s^\lambda = F_{0^-} - K + R$. Thus, upon maturity, the lender is paid back the face value K and immediately extends a new loan (with face value K), yielding additional payoff $F_{0^-} - K$ plus R ("origination fees") so that $F_s^\delta = K + F_{0^-} + R - K$. When default occurs, the lender is not paid back anything; with probability p^λ, the lender originates a new loan with face value K and accordingly earns $F_{0^-} - K + R$.

Thus, the continuation surplus at time t before maturity is characterized by

$$\begin{aligned} F_t &= \int_t^\infty e^{-(r+\delta)(s-t)-\int_t^s \lambda_u du}\left(1-\frac{\phi a_s^2}{2}-(\gamma-r)W_s+\delta F_s^\delta+p^\lambda\lambda_s F_s^\lambda\right)ds \\ &= \int_t^\infty e^{-(r+\delta)(s-t)-\int_t^s \lambda_u du}\left(1-\frac{\phi a_s^2}{2}-(\gamma-r)W_s+(\delta+p^\lambda\lambda_s)(F_{0^-}+R)-p^\lambda\lambda_s K\right)ds. \end{aligned}$$

Limited liability for investors requires $F_t - W_t \geq 0$ as well as $F_t^\lambda - dC_t^\lambda - W_t^\lambda = F_t^\lambda \geq 0$ and $F_t^\delta - dC_t^\delta - W_t^\delta \geq 0$.

As in our previous analysis, we conjecture and verify that all payoff-relevant quantities can be written as functions of $V = V_t$ only, in that $F_t = F(V)$, $W_t = W(V)$, $a_t = a(V)$, and $c_t = c(V)$. We omit time subscripts, unless necessary. Using above integral expression for F_t as well as invoking the dynamic programming principle, total surplus $F(V)$ satisfies the HJB equation:

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

subject to $W \in [0, F(V)]$ and $a = W/\phi$. Recall $F^\delta = F_{0^-} + R$ and $F^\lambda = F_{0^-} - K + R$.

The optimization with respect to monitoring effort a yields

$$a(V) = \frac{F(V) - p^\lambda F^\lambda - (\gamma - r)\phi - F'(V)[V+\phi]}{\phi}.$$

Finally, we characterize the boundary behavior of the HJB equation. For this sake, consider

$$V^B = \frac{W(V^B)}{\gamma+\delta+\lambda^B},$$

with $\lambda^B = \Lambda - q - a(V^B)$. The above HJB equation is solved for $V > V^B$ subject to the boundary condition

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

with

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

subject to $a = W/\phi$.

Optimal screening effort at time $t = 0^-$ is then determined according to:

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

where we assume that parameters imply $F_{0^-} > K$.

Notice that at maturity or default, a new loan is originated and the lender essentially solves the same problem as at time $t = 0^-$. As such, total surplus becomes F_{0^-} and the lender's continuation payoff is reset to W_0, so that $W_t^\delta = W_t^\lambda = W_0$ which pins down $dC_t^\lambda = -W_0$. As a next step, we determine dC_t^δ, i.e., the agent's payment upon maturity, and flow payouts c_t. Using $W_t^\lambda = W_t^\delta =$

$-dC_t^\lambda = W_0$, we can rewrite (B.71) to obtain

$$\hat{c}_t := c_t + \delta dC_t^\delta = (\gamma + \delta + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \delta W_0 - \dot{W}_t. \tag{B.72}$$

As the optimal contract pins down $W_t = W(V_t)$ and $\dot{W}_t = W'(V_t)\dot{V}_t$ as well as optimal monitoring a_t and default intensity λ_t, it also uniquely pins down $\hat{c}_t := c_t + \delta dC_t^\delta$. However, while $\hat{c}_t := c_t + \delta dC_t^\delta$ is uniquely pinned down, the individual components of this sum c_t and dC_t^δ are not unique and can be chosen arbitrarily, subject to $\hat{c}_t = c_t + \delta dC_t^\delta$ and $F_t^\delta - dC_t^\delta - W_t^\delta \geq 0$, i.e., $F_{0^-} + R - W_0 \geq dC_t^\delta$.

In the implementation below, we will pick a specific value of dC_t^δ, which then pins down uniquely the payout rate c_t via (B.72).

B.4.2 Implementation, Retention, and Commitment

The optimal contract in the model with repeated interactions leads to an optimal state-contingent level of continuation payoff for the agent $W_t = W(V_t)$, with dynamics W_t characterized in (B.71) or (B.72). We now discuss the implementation of the optimal contract by means of time-varying retention of a share by the agent. For this purpose, consider the market value of the loan

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

In optimum, the market value of the loan is a function of V_t, i.e., $L_t = L(V_t)$ and satisfies the ODE

$$(r + \lambda(V))L(V) = 1 + L'(V)\dot{V}$$

subject to $\lim_{V \to V^B} L(V) = \frac{1}{r+\Lambda-a(V^B)-q}$.

The lender retains a share $\beta_t = \beta(V_t)$. The lender's retention $\beta_t = \beta(V_t)$ is determined according to

$$c_t = \beta_t - \dot{\beta}_t L_t \quad \Longleftrightarrow \quad c(V_t) = \beta(V_t) - \beta'(V_t)\dot{V}_t L(V_t). \tag{B.73}$$

That is, the lender receives coupon payments at rate one per unit retained of the loan, i.e., in total $\beta(V_t)$ dollars, and obtains additional payoff $-\beta'(V_t)\dot{V}_t L(V_t)$ from selling the loan at market price $L(V_t)$. The sum of coupon payments and payoff from selling must equal contracted payouts in the implementation.

As a next step, we determine dC_t^δ, i.e., the agent's payment upon maturity. Using (B.73), we can rewrite (B.72):

$$\beta_t - \dot{\beta}_t L_t = (\gamma + \delta + \lambda_t)W_t + \frac{\phi a_t^2}{2} - \delta(W_0 + dC_t^\delta) - \dot{W}_t. \tag{B.74}$$

As a next step, we determine dC_t^δ, i.e., the agent's payment upon maturity.

Recall that we can pick arbitrary $dC_t^\delta \leq F_{0^-} + R - W_0$, which then pins down the payout rate c_t via (B.72) and accordingly $\beta_t - \dot{\beta}_t L_t$ via (B.74). We pick $dC_t^\delta = F_{0^-} - K - W_0 + \beta_t K + R = F_{0^-} + R - W_0 - (1 - \beta_t)K$. We motivate this choice as follows. First, on the maturity date t, the lender owns a fraction β_t of the maturing loan and thus is paid a fraction β_t of the face value K (i.e.,

$\beta_t K$). Second, the lender extracts the entire surplus $F_{0^-} - K + R$ from the follow-up origination. That is, $W_0^\delta + dC_t^\delta = F_{0^-} - (1-\beta_t)K + R$. As a result, using (B.74), we obtain the following ODE

$$\beta(V) - \beta'(V)\dot{V}L(V) = (\gamma+\delta+\lambda(V))W(V) + \frac{\phi a(V)^2}{2} - \delta(F_{0^-} + R - (1-\beta(V))K) - W'(V)\dot{V}, \quad \text{(B.75)}$$

which determines the (state-contingent) retention level $\beta(V)$. The ODE is solved for $V > V^B$ with boundary condition

$$\lim_{V \to V^B} \beta(V) = \frac{1}{1+\delta K}\left((\gamma+\delta+\lambda^B)W(V^B) + \frac{\phi a(V^B)^2}{2} - \delta(F_{0^-} + R - K)\right) \quad \text{(B.76)}$$

We assume that a solution exists and is unique.

According to (B.75) and (B.76), expected future payoffs from loan origination, as captured by $\delta(F_{0^-} + R - (1-\beta_t)K)$, *all else equal* reduce flow payouts to the agent c_t and thus the required retention level β_t. This effect is stronger when the loan maturity $1/\delta$ is shorter (i.e., when δ is larger) as the payoff from repeated origination only materializes when the outstanding loan matures. Likewise, the expected future payoff from origination is larger and thus the required retention level is lower when R (capturing origination fees) or $F_{0^-} - K$ (capturing "net" surplus from origination) are larger. That is, expected future payoffs from loan origination provide screening and monitoring incentives to the lender and are a substitute to loan retention in incentive provision.

We now discuss whether lender commitment is necessary for the implementation. For this purpose, consider the extreme case in which the lender cannot commit to retaining part of the loan. However, if the lender deviates from the retention path stipulated in the contract at any point in time, it can no longer sell loans to investors and, as a result, originate loans (due to the fact that other lenders can better price loans that they can resell to investors). The underlying assumption is that in the absence of investors who buy the loans originated by the lender, the lender is not able or willing to originate loans in autarky, e.g., because it is not profitable to do so if it cannot sell the loans or because it is not able to originate loans on its own due to capital constraints. In other words, investors cut the relationship and thus play a grim-trigger strategy. The market price at time t equals L_t. If the lender is able to sell its entire stake at price L_t (i.e., there is no price impact), it earns $L_t\beta_t$ but looses its continuation payoff W_t. Thus, the lender prefers not to sell its entire stake if $L_t\beta_t \leq W_t$ or in terms of the state variable

$$L(V)\beta(V) \leq W(V) \quad \text{for all} \quad V \in (V^B, V_0) \quad \text{(B.77)}$$

holds. Given the grim-trigger strategy of investors, selling the entire stake at market price $L(V_t) = L_t$ (i.e. without price impact) is the best possible scenario for the agent. Thus, (B.77) is sufficient for the implementation to work even in the lack of commitment.

We now numerically check under what circumstances condition (B.77) holds. For this purpose, we set the average maturity of the loan to 4.43 years, as reported in Blickle et al. (2022), by setting $\delta \approx 0.23$. We further set $K = 0.75$, normalize $R = p^\lambda = 0$, and use our baseline parameters otherwise. We numerically solve the model with repeated interactions as well as the implementation of the optimal contract for a wide range of parameters. Figure B.4 performs comparative static analysis with respect to R (origination fees), $1/\delta$ (maturity), and K (face value) and plots

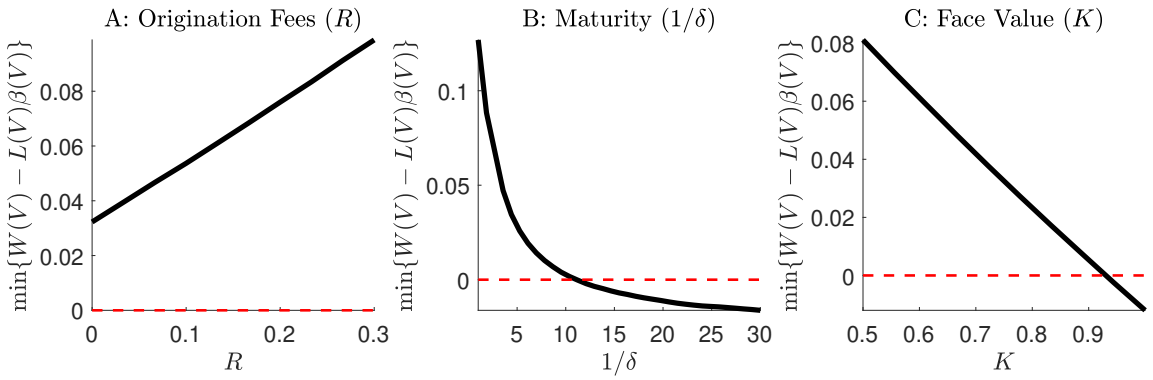

Figure B.4: **Comparative Statics and Loan Portfolios.** This figure plots $\min\{W(V) - L(V)\beta(V)\} := \min\{W(V) - L(V)\beta(V) : V \in [V^B(q), V_0]$ against R, $1/\delta$, and K under our baseline parameters as well as $\delta = 0.23$, $K = 0.75$, and $R = p^\lambda = 0$. Notice that (B.77), so that the lender would not like to deviate from the stipulated retention path, holds if and only $\min\{W(V) - L(V)\beta(V)\} \geq 0$

$\min\{W(V) - L(V)\beta(V)\} := \min\{W(V) - L(V)\beta(V) : V \in [V^B(q), V_0]\}$ against R (Panel A), $1/\delta$ (Panel B), and K (Panel C). Notice that (B.77) holds if and only if $\min\{W(V) - L(V)\beta(V)\} \geq 0$ in which case the lender does not benefit from deviating from the stipulated retention path. Notice that the lender indeed does not benefit from deviating and (B.77) holds when expected future payoffs from loan origination, as captured by $\delta(F_{0^-} + R - (1 - \beta_t)K)$ are large. Thus, (B.77) holds when the loan maturity $1/\delta$ is not too long (e.g., smaller than 10 years in our numerical example) and the face value K is not too large. Notably, (B.77) holds in our baseline parameters, in addition to $1/\delta = 4.43$, $K = 0.75$, and $R = 0$ and, as is observed, across a wide range of parameters even when $R = 0$.

We emphasize that (B.77) should be interpreted as a sufficient condition for the implementation to be robust to commitment problems. Both in practice as well as in theory, loan sales would have a price impact rendering large discrete sales not attractive for the lender and leading the lender to retain its stake in the loan, effectively allowing the lender to commit to some retention path. Such price impact could arise for example because loan sales reduce lender incentives to monitor and thus the loan's value. Second, depending on the model interpretation and application, the market for loans might be illiquid, especially shortly after origination, which allows the lender to commit to retention during this period. Specifically, if we interpret time $t = 0$ as the beginning of the primary market within syndicated lending (for details on the syndication process, see Bruche et al. (2020)) while the secondary market opens at some time $T > 0$, then the lender cannot easily sell its stake in the primary market over $[0, T]$, again giving the lender some commitment power. Third, in practice within syndicated lending, loan sales (e.g., to institutional investors) are often pre-committed at origination. That is, at origination, the lender commits to sell a certain share to investors at a later time. As such, there is some form of commitment to loan sales in the market for syndicated loans, supporting the plausibility and empirical relevance of our results.

B.5 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 $a_t = \hat{a}_t$.

B.5.1 Model 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 $dC_t^{\delta,s}$ ($dC_t^{\delta,m}$) dollars.

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

$$dW_t^s = (\gamma + \lambda_t + \delta)W_t^s dt - dC_t^s - \delta dC_t^{s,\delta} dt \tag{B.78}$$

$$dW_t^m = (\gamma + \lambda_t + \delta)W_t^m dt - dC_t^m + \frac{\phi a_t^2}{2} dt - \delta dC_t^{m,\delta} dt. \tag{B.79}$$

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

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

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 (B.78) with respect to q to obtain for the screener's incentives $V_t := \frac{\partial W_t^s}{\partial q}$:

$$dV_t = (\gamma + \lambda_t + \delta) V_t dt - W_t^s dt. \tag{B.80}$$

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 (A.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

$$\begin{aligned}(r+\delta)F(V) = \max_{a, W^m, W^s} \Big\{ & 1 - \frac{\phi a^2}{2} - (\gamma - r)(W^m + W^s) - \lambda F(V) \\ & + \delta F^\delta + F'(V)\big((\gamma + \lambda + \delta)V - W^s\big)\Big\}. \end{aligned} \tag{B.81}$$

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 (B.81) (subject to a boundary condition specified below).

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

$$W^s(V)\begin{cases}=0 & \text{if } F'(V)>-(\gamma-r)\\ \in[0,F(V)-W^m(V)] & \text{if } F'(V)=-(\gamma-r)\\ =F(V)-W^m(V) & \text{if } F'(V)<-(\gamma-r).\end{cases} \tag{B.82}$$

As in the baseline, it follows that $\lim_{t\to\infty} V_t = V^B(q)$, where $V^B(q)$ is the level of screening incentives in the benchmark without screening moral hazard (given q).[39] It follows that $V^B(q)=0$, as absent screening moral hazard it is optimal to set $V_t = W^s_t = 0$ at all times $t \geq 0$.

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

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

Owing to (B.82), 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 (B.81) and simplifying leads to the ordinary differential equation

$$(\gamma+\delta)F(V) = \max_{a,W^m}\Big\{1-\frac{\phi a^2}{2}-\lambda F(V)+\delta F^\delta + F'(V)\big((\gamma+\lambda+\delta)V-F(V)+W^m\big)\Big\}, \tag{B.83}$$

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

As in the main text (compare Section 4.3), we consider $F^\delta = F(V)$, so (B.83) simplifies to

$$\gamma F(V) = \max_{a,W^m}\Big\{1-\frac{\phi a^2}{2}-\lambda F(V)+F'(V)\big((\gamma+\lambda+\delta)V-F(V)+W^m\big)\Big\}. \tag{B.84}$$

Using the envelope theorem to totally differentiate the HJB equation (B.84) (under the optimal control $W^m = \phi a$) with respect to V yields

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

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 (B.80)). 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 (B.83) with respect to monitoring effort yields

$$a(V) = \frac{F(V)-F'(V)V+F'(V)\phi}{\phi}. \tag{B.85}$$

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

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 (B.82) that $F'(0) \geq -(\gamma - r)$. As $F'(V) < -(\gamma - r)$ for all $V > 0$, it follows — by continuity of $F'(V)$ — that $\lim_{V \to 0} F'(V) = -(\gamma - r)$. An alternative way to derive this boundary condition is as follows. Comparing (17) with (B.83), one can see that

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

is equivalent to

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

which is then natural the boundary condition for the ODE (B.83) as V approaches zero. We assume that a unique solution to (B.83) (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 B.3, 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 $V_0 = \kappa q$.

B.5.2 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 (B.85), 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 $V_t < V_0 \leq \bar{\kappa} q$. 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., $\dot{a}_t = a'(V_t)\dot{V}_t > 0$.

B.5.3 Analysis

We now analyze how bundling screening and monitoring tasks (as opposed to separating them) changes lender's incentives and screening/monitoring efforts, total surplus, and credit risk. For this sake, Figure B.5 plots the percentage change of initial monitoring a_0 (first row), optimal monitoring $q = q^*$ (second row), total initial surplus F_{0^-} (third row), and expected time to default at $t = 0$ (fourth row) upon bundling against κ, ϕ, Λ, and $1/\delta$. Notice that bundling monitoring and screening leads to positive synergies, while 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 B.5 illustrates these findings and shows that they are robust to changes in the κ, ϕ, Λ, and $1/\delta$. Under all parameters considered, bundling increases (initial) monitoring (i.e., $\Delta a_0 > 0$), screening ($\Delta q^* > 0$), and total surplus ($\Delta F_{0^-} > 0$). 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 B.5, 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.

B.6 Micro-foundation of Baseline Assumptions

This section provides a detailed micro-foundation for our reduced form modeling of the lender's screening and its impact on the default rate λ_t. Unless otherwise stated, we maintain the assumptions of the baseline model. We consider an environment with two types $x \in \{G, B\}$ of potential borrowers, a good (G) borrower with low default risk and a bad borrower (B) with higher default risk. There is a large mass of borrowers who would like to borrow from the lender at time 0. The ex-ante proportion of good type borrowers is $\omega \in [0, 1]$. Thus, if the lender encounters a randomly drawn borrower, this borrower is good with probability ω and bad with probability $1 - \omega$. If the lender extends a loan to the borrower of type x, this loan pays coupon at a rate normalized to one up to default.

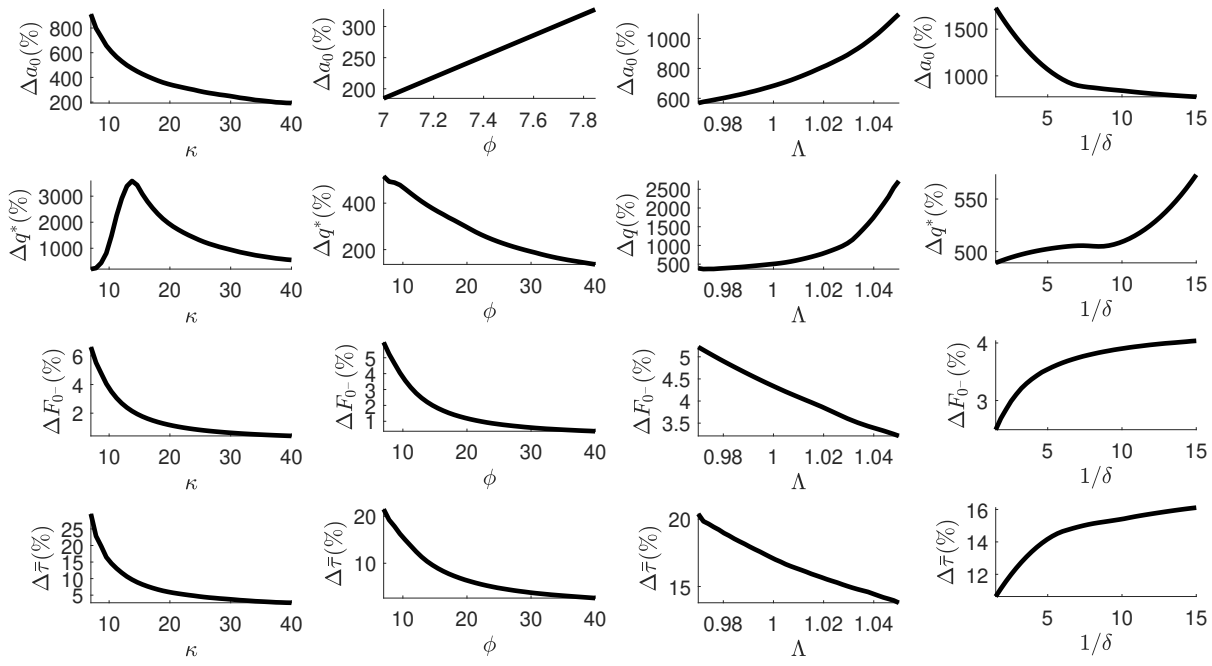

Figure B.5: **The effects of bundling screening and monitoring.** Δa_0 denotes the percentage change in monitoring effort at $t = 0$ due to bundling. Δq^* denotes the percentage change in screening effort due to bundling. ΔF_{0^-} denotes the percentage change in total surplus at $t = 0^-$ caused by bundling. $\Delta \overline{\tau}$ 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.

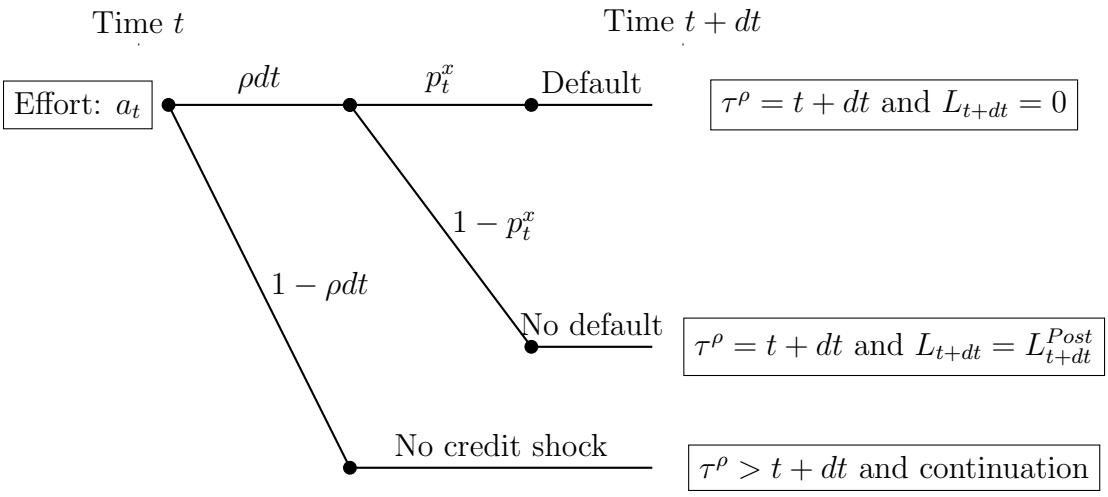

Figure B.6: Heuristic timing over $[t, t+dt)$. The branches of the tree contain the probabilities of the respective random event over $[t, t + dt)$.

B.6.1 Default Risk

For tractability, we model uncertainty as in, e.g., Board and Meyer-ter Vehn (2013), Hoffmann and Pfeil (2021), Gryglewicz, Mayer, and Morellec (2021), Mayer (2022), and Hu and Varas (2021a,b). Specifically, default only occurs upon an exogenous and publicly observable "credit shock" which arrives at constant, exogenous intensity $\rho > 0$. For simplicity, there is maximally one credit shock which arrives at random time τ^ρ; after time τ^ρ, there is no more uncertainty.[40] If this credit shock hits at time t, a type-x borrower defaults with probability p^x. The credit shock can be interpreted as an aggregate negative shock to the economy (such as a financial crisis), an aggregate earnings shock (such as Covid-19), or an aggregate cost of financing shock (such as an increase in interest rates as loans are generally floating rate instruments). Good firms are more likely able to withstand this shock than bad firms, in that the probability of default for a type-x borrower over any $[t, t + dt)$ is $p_t^x \rho dt$ with $p_t^G < p_t^B$. If the borrower defaults upon credit shock, the recovery value is for simplicity zero and the game ends. If the borrower survives the credit shock (with probability $1 - p_t^x$), the market value of the loan becomes L_t^{Post} (characterized below). Figure B.6 depicts the heuristic timing over a short time period $[t, t + dt)$.

In the following, we assume that

$$p_t^G = \overline{p} - \chi_G a_t \quad \text{and} \quad p_t^B = 1 - \chi_B a_t$$

for some constant $\chi_x \geq 0$ and $\overline{p} < 1$. That is, we assume as in the baseline that the lender's monitoring effort a_t (chosen before the realization of the credit shock) reduces the probability of default upon credit shock. Depending on the exact values of χ_g and χ_b, monitoring can have different impact in reducing default risk for good and bad borrowers. We assume that $\overline{p} - \chi_g \bar{a} < 1 - \chi_b \bar{a}$, i.e., even if monitoring effort a_t is at its maximum $\bar{a}$, the bad type borrower has higher default risk.

[40] The expected time to τ^ρ then equals $1/\rho$.

Last, if the lender believes that the borrower is good, i.e., $x = G$, with probability $\hat{\omega}_t$, then the expected probability of default over $[t, t + dt)$ from the lender's point of view is $\lambda_t dt$ with

$$\lambda_t = \rho\big[\hat{\omega}_t p_t^G + (1 - \hat{\omega}_t) p_t^B\big].$$

As we show next, $\hat{\omega}_t$ depends on the lender's screening effort q and takes the form $\hat{\omega}_t = \hat{\omega}(q)$.

B.6.2 Loan Origination and Screening

The sequence of events at time $t = 0^-$ (i.e., just before screening) is as follows. First, the lender and investors sign a contract $\mathcal{C} = \{dC_t, \hat{a}_t, \hat{q}\}$, stipulating incremental payouts to the lender dC_t as well as recommended screening effort $\hat{q}$ and monitoring efforts $(\hat{a}_t)_{t\geq 0}$. Second, the lender is randomly matched with a borrower and receives a signal S about borrower type; the signal's precision depends on the lender's screening effort. Third, the lender either accepts the borrower, in which case the loan to the borrower is originated, or rejects the borrower, in which case the lender is randomly matched with another borrower and we are back in the second step.

The signal S about borrower's type x can take two values g and b, where $Pr(S = b|x = B) = 1 - f(q)$ and $Pr(S = b|x = G) = 0$. That is, the signal S may only yield false positives, i.e., it may yield g for a bad borrower. On the other hand, if the signal takes value b, a bad type borrower is identified (i.e., $Pr(s = b|x = G) = 0$). The probability of a false-positive signal reads $Pr(s = g|x = B) = f(q)$, where $f(q)$ decreases with q (subject to standard regularity requirements, such as $f(q) \in [0, 1]$ for $q \in [0, \bar{q}]$). In other words, the precision of the signal, that is, $1 - f(q)$, increases with the lender's screening effort, so that higher screening effort q allows the lender to better differentiate good from bad borrowers. By Bayes' rule, observing $S = g$, the lender believes that the borrower is good with probability

$$\hat{\omega}(q) := Pr(x = G|S = g) = \frac{Pr(x = G \ \wedge \ S = g)}{Pr(S = g)} = \frac{\omega}{\omega + (1 - \omega) f(q)}, \tag{B.86}$$

so that $\hat{\omega}'(q) > 0$.

We assume that the lender rejects the borrower if it observes $S = b$ in which case a bad type borrower is identified and the lender moves on and screens the next borrower. As such, the lender searches for a borrower until a good signal $S = g$ is observed. Then, the lender originates the loan to the borrower, so the lender and investors receive coupon payments at rate one from time $t = 0$ onwards whereby the lender is paid dC_t per dt.

Importantly, after time $t = 0$, there is no more learning about borrower type unless there is a credit shock (which, as we recall, is publicly observable). The only reason behind the modeling of default only occurring at publicly observable credit shocks is to preclude dynamic learning about borrower type over time.[41] In Section B.6.10, we extend the analysis to multiple credit shocks in which case there is learning and belief updating if the loan survives a credit shock. In this extension, the lender's belief about borrower type moves over time and becomes an additional state variable, thereby significantly complicating the analysis without delivering new economic insights.

[41] Analogously, Hu and Varas (2021b) also add tractability to their setting by employing similar modeling of default/failure.

B.6.3 Screening and Monitoring: Substitutes vs. Complements

Recall that at any point in time t, the (expected) default intensity from the lender's and investors' point of view is:

$$\lambda_t = \rho\big[\hat{\omega}(q)p_t^G + (1-\hat{\omega}(q))p_t^B\big].$$

We assume the following tractable functional form of $f(q)$:

$$f(q) = \frac{\omega(1-\zeta q)}{(1-\omega)\zeta q} \quad \Longleftrightarrow \quad \hat{\omega}(q) = \zeta q,$$

where we assume the parameters $\zeta, \bar{q}$, and ω are such that $\hat{\omega}(q)$ and $f(q)$ are well-behaved and lie between zero and one. We can generally write using $\hat{\omega}(q) = \zeta q$:

$$\begin{aligned}\lambda_t &= \rho - \rho\zeta(1-\overline{p})q - a_t\chi_B - a_t q\zeta(\chi_G - \chi_B)\\ &=: \rho - \alpha_q q - \alpha_a a_t - \alpha_{aq} q a_t\end{aligned}$$

with $\alpha_q = \rho\zeta(1-\overline{p})$, $\alpha_a = \chi_B$, and $\alpha_{aq} = \zeta(\chi_G - \chi_B)$. Notably, our baseline specification with default intensity (1) is obtained upon setting $\rho = \Lambda$, $\alpha_a = \alpha_q = 1$, and $\alpha_{aq} = 0$ (i.e., $\chi_G = \chi_B$).

Depending on χ_G and χ_B, screening and monitoring can be complements or substitutes in reducing default risk. When $\chi_G > \chi_B$, monitoring has a larger effect on reducing default risk for good types than for bad types. We then have $\frac{\partial^2\lambda_t}{\partial a_t \partial q} = -\zeta(\chi_G - \chi_B) < 0$ and screening and monitoring are complements in reducing default risk. When $\chi_G < \chi_B$, monitoring has a smaller effect on reducing default risk for good types than for bad types. In this case, we have $\frac{\partial^2\lambda_t}{\partial a_t \partial q} = -\zeta(\chi_G - \chi_B) > 0$ and screening and monitoring are substitutes in reducing default risk. When $\chi_G = \chi_B$, they are neither complements nor substitutes, as we assume in the baseline.

It is ex-ante unclear whether $\chi_G > \chi_B$ or $\chi_G < \chi_B$ prevails in credit markets and this might depend on the specific market one is analyzing. For instance, when good borrowers are perfectly safe regardless of monitoring, we would expect $\chi_G < \chi_B$, and screening and monitoring are substitutes. On the other hand, when bad borrowers are very risky and default with a very high probability regardless of monitoring, we get $\chi_G > \chi_B$ and screening and monitoring would be complements. Crucially, and as we show, the model's main results obtain regardless of whether we assume that screening and monitoring are substitutes or complements (see Section 2.2.4 in the main text).

In our baseline model, we assume that screening and monitoring are neither substitutes nor complements for several reasons. First, we think it is unclear whether $\chi_G > \chi_B$ or $\chi_G < \chi_B$ prevails in practice, so we do not want to take a stance. Second, by not making assumptions on substitutability or complementarity, we afford maximum theoretical clarity, tractability, and reduce the number of model parameters to focus on the paper's implications on lender incentives and loan sale and retention dynamics. Third, we show that for pure incentive provision, screening and monitoring have the tendency to behave like complements and we rather focus on the endogenous than the assumed relationship.

B.6.4 Dynamic Optimization and Contracting

We now provide the heuristic solution to the dynamic contracting problem and show that, under certain assumptions, it becomes isomorphic to that in the baseline model. As in the baseline model, we start our analysis by taking the level of screening q as given.

Recall that, after the loan is originated, the lender faces a good borrower with probability $\hat{\omega}(q)$, pinning down the expected rate of default as well as the expected probability of default

$$p_t := \left[\hat{\omega}(q)p_t^G + (1-\hat{\omega}(q))p_t^B\right] = 1 - \frac{\alpha_q}{\rho}q - \frac{\alpha_a}{\rho}a_t - \frac{\alpha_{aq}}{\rho}qa_t. \tag{B.87}$$

We normalize $\alpha_q = \alpha_a = 1$, set $\rho = \Lambda$, and set $\alpha = \alpha_{aq}$ so that $\lambda_t = \rho p_t = \Lambda - a_t - q - \alpha a_t q$, yielding a similar "default intensity" λ_t as in Section 2.2.4 and our baseline when $\alpha = 0$.

Given screening and monitoring efforts, the lender's continuation payoff at any time t, before the credit shock, reads

$$W_t = \int_t^\infty e^{-(\gamma+\rho)(s-t)}\left(c_s - \frac{\phi a_t^2}{2} + \rho(1-p_s)W_s^{Post}\right), \tag{B.88}$$

where W_s^{Post} is the lender's continuation payoff if a credit shock hits at time s and the loan does *not* default (characterized below). We again may conjecture and verify that for $t \in (0, \tau^\rho)$ payments are smooth, i.e., $dC_t = c_t dt$. Upon default, the lender loses its entire continuation payoff (stake). Differentiation with respect to time t yields

$$\begin{aligned}\dot{W}_t &= (\gamma+\rho)W_t + \frac{\phi a_t^2}{2} - c_t - \rho(1-p_t)W_t^{Post} \\ &= (\gamma+\rho)W_t + \frac{\phi a_t^2}{2} - c_t - (a_t + q + \alpha a_t q)W_t^{Post}.\end{aligned} \tag{B.89}$$

By the dynamic programming principle, the lender chooses monitoring effort a_t according to:

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

so that (provided $a_t \in [0, \bar{a}]$):

$$a_t = \frac{(1+\alpha q)W_t^{Post}}{\phi}. \tag{B.90}$$

The investors' payoff is given by

$$P_t = \int_t^\infty e^{-(r+\rho)(s-t)}\left(c_s - \frac{\phi a_t^2}{2} + \rho(1-p_s)P_s^{Post}\right), \tag{B.91}$$

where P_s^{Post} is investor payoff upon surviving credit shock at s.

Next, we characterize screening incentives. As in the baseline analysis, we calculate the law of motion of $V_t := \frac{\partial W_t}{\partial q}$ by totally differentiating (B.89) with respect to q:

$$\dot{V}_t = (\gamma+\rho)V_t - (1+\alpha a_t)W_t^{Post} - (a_t + q + \alpha a_t q)V_t^{Post},$$

with $V_t^{Post} := \frac{\partial W_t^{Post}}{\partial q}$. That is,

$$V_t = \int_t^\infty e^{-(\gamma+\rho)(s-t)}\Big((1+\alpha a_s)W_s^{Post} + (a_s + q + \alpha a_s q)V_s^{Post}\Big)ds.$$

As there is only one credit shock, the impact of screening effort only lasts until the credit shock (which occurs at time τ^ρ), so that $V_t^{Post} = 0$.

Next, and again as in the baseline model, total surplus–split between the lender and investors–from time t (before the credit shock) reads

$$F_t = \int_t^\infty e^{-(r+\rho)(s-t)}\left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s + \rho(1-p_s)F_s^{Post}\right)ds, \tag{B.92}$$

where F_s^{Post} is the continuation surplus at s if the loan survives a credit shock at s.

Total surplus is split between the lender and investors. So, before the credit shock, investors' stake/value is $P_t = F_t - W_t$ (that is, $F_t = P_t + W_t$) and, upon surviving the credit shock at t, investors' value is $P_t^{Post} = F_t^{Post} - W_t^{Post}$.

Surviving the credit shock is good news, so naturally the loan value as well as total surplus should go up in this event. Therefore, we generally consider $F_t^{Post} \geq F_t$; we specify F_t^{Post} later and show how specific assumptions on F_t^{Post} make the model analogous to the baseline. It is then also natural to consider that both investors and lender benefit from this positive outcome. Specifically, we impose a monotonicity requirement on the lender-investor contract, in that neither lender nor investor can be worse off following the credit shock if the firm survives. That is, we impose $W_t^{Post} \geq W_t$ as well as

$$P_t^{Post} \geq P_t \quad \Longleftrightarrow \quad W_t^{Post} \leq F_t^{Post} - F_t + W_t. \tag{B.93}$$

Again, the monotonicity assumption is natural. In practice and in our baseline implementation of the contract, investors and lender both hold shares of the loan. As such, they should both (at least weakly) benefit when loan value increases.

B.6.5 HJB Equation

Next, as in the baseline model, W_t and W_t^{Post} become control variables while V_t is the state variable in the dynamic optimization problem. We omit time subscripts, unless necessary. As such, total surplus is a function of V_t only, i.e., $F_t = F(V_t)$. In state $V_t = V$, the HJB equation characterizing $F(V)$ reads (applying standard arguments and using the integral expression for total continuation surplus in (B.92):

$$\begin{aligned}(r+\rho)F(V) = \max_{a_t\in[0,\bar{a}],W,W^{Post}} \Bigg\{ & 1 - \frac{\phi a^2}{2} - (\gamma - r)W + (a + q + \alpha a q)F^{Post} \\ & + F'(V)\Big((\gamma+\rho)V - (1+\alpha a)W^{Post} - (a + q + \alpha a q)V^{Post}\Big)\Bigg\},\end{aligned}$$

which is solved subject to (B.90)–i.e., $a = W^{Post}/\phi$–(B.93)–i.e., $W \geq F(V) + W^{Post} - F^{Post}$–$W^{Post} \geq W$, and $W \geq 0$. Notice that the right-hand side decreases with W and a does not depend

on W. Thus, holding other controls equal, it is optimal to minimize W subject to the constraint (B.93), i.e., $W \geq F(V) + W^{Post} - F^{Post}$, and $W \geq 0$. As a result, $W = \max\{F(V) + W^{Post} - F^{Post}, 0\}$. Also, recall that $V^{Post} = \frac{\partial W^{Post}}{\partial q} = 0$, because the impact of screening lasts only until the first and only credit shock in that W^{Post} does not depend on q.

We use (B.90) and $W = \max\{F(V) + W^{Post} - F^{Post}, 0\} =: (F(V) + W^{Post} - F^{Post})^+$, where $(y)^+ = \max\{y, 0\}$, to rewrite above HJB equation as

$$(r+\rho)F(V) = \max_{W^{Post} \in [0, \bar{a}\phi]} \left\{ 1 - \frac{(W^{Post})^2}{2\phi} - (\gamma - r)(F(V) + W^{Post} - F^{Post})^+ \right.$$
$$\left. + \left(\frac{W^{Post}}{\phi} + q + \frac{\alpha q W^{Post}}{\phi} \right) F^{Post} + F'(V) \left[(\gamma+\rho)V - \left(1 + \frac{\alpha W^{Post}}{\phi} \right) W^{Post} \right] \right\}, \quad \text{(B.94)}$$

with $W^{Post} = W + F^{Post} - F(V)$.

Boundary conditions. As in the baseline model, we expect $\lim_{t\to\infty} V_t = V^B$ and $\lim_{t\to\infty} \dot{V}_t = 0$. That is,

$$V^B = \frac{W^{B,Post}(1 + \alpha a^B)}{\gamma + \rho}.$$

Notice that $a^B = W^{B,Post}/\phi$ and $W^{B,Post}$ are determined according to the optimization

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

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

which follows from (B.94) evaluated at $V = V^B$ and $F(V^B) = F^B$ after setting $\dot{V} = 0$.

B.6.6 Characterizing F^{Post} and Connection to the Baseline Model

We have so far not specified the continuation surplus upon credit shock and no default, i.e., F^{Post}. The exact value of F^{Post} does not affect the model's key qualitative implications. The model greatly simplifies upon assuming $F^{Post} = F(V)$ in which case $W = W^{Post}$ using (B.93) and we can rewrite the HJB equation as

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

with $\lambda = \Lambda - a - q - \alpha a q$. This HJB equation is analogous and similar to the baseline HJB equation (23). It in fact becomes identical upon setting $\alpha = 0$, i.e., upon assuming $\lambda_t = \Lambda - a_t - q$ from (1).

The assumption $F^{Post} = F(V)$ can be interpreted as follows: Upon the credit shock, the loan either defaults or matures. In case of no default and maturity, the loan pays back the face value. As in Section 4.3, we assume value matching, i.e., upon maturity the lender-investor surplus remains

unchanged as is captured by $F(V) = F^{Post}$. Alternatively, we can assume that the loan is simply sold at price L at time τ^ρ where we expect the exact value of L not to affect the model's qualitative implications. That is, $F^{Post} = L$.[42]

B.6.7 Optimal Choice of Screening Effort

We now characterize the optimal choice of screening effort as well as the corresponding incentive condition. Recall that $t = 0^-$ denotes the time just before screening effort is chosen while $t = 0$ denotes the time just after screening effort is chosen. At time $t = 0^-$, the lender maximizes:

$$\max_{q\in[0,\bar{q}]} \big(\omega + (1-\omega)f(q)\big)\big(P_0 + W_0 - K\big) + \Big[1 - \big(\omega + (1-\omega)f(q)\big)\Big]F_B - \frac{\kappa q^2}{2},$$

where F_B is the surplus upon receiving a bad signal and K is the amount borrowed by the borrower (i.e., the loan size) which is assumed not to depend on hidden (non-contractible) screening effort q. To understand the above objective, note that when the lender exerts screening effort q, the signal is good with probability $\omega + (1-\omega)f(q)$ in which case the lender originates a loan and extracts (post-screening) surplus $F_0 - K = P_0 + W_0 - K$ from origination (after the lent amount K is paid to the borrower at $t = 0$, the continuation surplus becomes F_0). Otherwise, with probability $1-\omega-(1-\omega)f(q)$, the lender receives a bad signal and rejects the borrower, and the payoff becomes F_B.[43]

Note that P_0 does not depend on actual q, while W_0 does. Likewise, F_B and K do not depend on the screening effort q. Provided that the first order approach is valid, the incentive condition regarding screening becomes with $V_t = \frac{\partial W_t}{\partial q}$:

$$\big(\omega + (1-\omega)f(q)\big)V_0 - (1-\omega)f'(q)\big[F_0 - F_B - K\big] = \kappa q.$$

At time $t = 0^-$, respecting the incentive condition for screening, total surplus reads

$$F_0^- = \big(\omega + (1-\omega)f(q)\big)(F_0 - K) + \Big[1 - \big(\omega + (1-\omega)f(q)\big)\Big]F_B - \frac{\kappa q^2}{2},$$

with $F_0 - K$ being the post screening surplus from origination. After the lent amount K is paid to the borrower at $t = 0$, the continuation surplus becomes F_0. Because the borrower is rejected upon bad signal S, we have $F_B = F_{0^-}$ and we obtain

$$F_0^- = \big(\omega + (1-\omega)f(q)\big)(F_0 - K) + \big[1 - \big(\omega + (1-\omega)f(q)\big)\big]F_{0^-} - \frac{\kappa q^2}{2}$$

so that

$$F_0^- = F_0 - \frac{\kappa q^2}{2\big(\omega + (1-\omega)f(q)\big)} - K,$$

[42] For instance, we could assume that the value after the credit shock equals the riskfree value $L = 1/r$

[43] In principle, rejection of a borrower as well as acceptance of a borrower is contractible. If the agent were to receive some positive payments to rejection, it would just easily collect them by rejecting. Thus, there cannot be positive payments for rejection. A punishment for rejection would dis-incentivize the agent to screen and thus is likely not optimal either. We do not formally consider this case.

i.e., $F_0^- = F_0 - \kappa(q)$ with a cost function $\kappa(q)$.

Using $F_0^- = F_B$, the incentive condition then becomes

$$V_0 = \frac{1}{\omega + (1-\omega)f(q)} \left(\kappa q + \frac{(1-\omega)f'(q)\kappa q^2}{2\big(\omega + (1-\omega)f(q)\big)} \right)$$

As such, the incentive condition for screening becomes significantly more complicated relative to the baseline, but the key economic implications are the same: The state variable V_t encapsulates screening incentives and screening incentives at $t = 0$ translate into higher screening effort, while incentivizing high screening q becomes harder when κ is larger.

B.6.8 Loan Size and Yield

The above analysis goes through for any loan size K as long as the lender's participation constraint $F_0^- \geq 0$ is satisfied. More specifically, as we argue in what follows, the exact value of K does not change the contract dynamics qualitatively, and merely affects the initial surplus.

We now endogenize the loan size K. To this end, we fix the coupon payments at one dollar, so that the yield on the loan becomes $R^* = 1/K$. Assume the borrower has a project that pays 1 up to default, but it only derives utility from immediate consumption at time $t = 0^-$; the project can be sold for $\overline{B}$ dollars which constitutes the agent's outside option.

Consider that at time $t = 0^-$ (i.e., just before screening), the borrower and the lender are matched and bargain over the loan's size K (given the coupon payments of one dollar per unit of time), in case the loan is eventually originated. When bargaining, the lender and the borrower take as given that the borrower is accepted and the loan is originated at $t = 0$ if and only if the signal during the screening process is good. The borrower has bargaining power measured by $\alpha \in [0, 1]$ and outside option $\overline{B}$ and the lender has bargaining power $1 - \alpha$ and (exogenous) outside option $\overline{F}$.

The additional surplus generated from the lender-borrower interaction at $t = 0^-$ is then:

$$F_{0^-} - \overline{B} - \overline{F}.$$

We assume that the lender extracts its outside option plus fraction a α of this surplus, so that[44]

$$K = \overline{B} + \alpha(F_{0^-} - \overline{B} - \overline{F})$$

and

$$R^* = \frac{1}{\overline{B} + \alpha(F_{0^-} - \overline{B} - \overline{F})}.$$

Notice that the yield depends on the expected borrower and lender fundamentals (e.g., the cost of monitoring and screening), which in turn affect optimal screening, via F_{0^-}. Generally, we expect that $\frac{\partial F_{0^-}}{\partial \kappa} < 0$ and $\frac{\partial q^*}{\partial \kappa} < 0$, so that $\frac{\partial R^*}{\partial q} > 0$ and lower screening is associated with a higher yield on the loan.

[44] One could micro-found this assumption by considering Nash bargaining between the lender and the borrower.

The lender's payoff at $t = 0^-$ is

$$F_{0^-} - K = \overline{F} + (1-\alpha)(F_{0^-} - \overline{B} - \overline{F}) = \alpha\overline{F} + (1-\alpha)(F_{0^-} - \overline{B}).$$

We assume that $\overline{B}$ is not prohibitively large so that loan origination is optimal, i.e., $F_{0^-} > \overline{B}$. Importantly, the optimal lender-investor contract is chosen to maximize F_{0^-}. The shape of this contract as well as its dynamics do not depend on the exact value of K (or its determinants, e.g., α or $\overline{B}$).

B.6.9 Credit Ratings

We can also examine the effects of ratings on outcome variables in this model variant. The credit rating will be a publicly observable signal R informative about borrower type, taking two values g and b. In line with the screening signal structure, assume there are only false positives, i.e., $Pr(R = b|G) = 0$ while $Pr(R = g|B) = 1 - \sigma$, where $\sigma \in [0, 1]$ is the precision of the rating. For simplicity, consider limit case $\sigma = 1$ of a perfectly informative rating.

The timing of events at origination is as follows. First, at time $t = 0^-$, the lender-investor contract $\mathcal{C}$ is signed. Second, also at time $t = 0^-$, the lender chooses screening effort q, observes a signal $S \in \{b, g\}$ about borrower type $x \in \{B, G\}$, and accepts the borrower if and only if $S = g$. Third, immediately after the loan is originated and at time $t = 0$ (after screening), the credit rating $R \in \{b, g\}$ is publicly observed. Fourth, the initial transfer $dC_0(R)$ takes place and can be contingent on R. Further, the contract $\mathcal{C}$ and in particular the continuation contract from time $t = 0$ onward can also be contingent on credit rating R, i.e., the rating R is contractible.

As such, continuation surplus and continuation payoff for the agent at time $t = 0$ can be contingent on R and, because the rating is perfectly informative, contingent on borrower type x, so that we can write $F_0 = F_0(x)$ and $W_0 = W_0(x)$. Limited liability for lender and investors requires at $t = 0$ and just after the rating that:

$$\begin{aligned} dC(x) + W_0(x) &\geq 0 \\ dC(x) &\leq F_0(x) - W_0(x). \end{aligned}$$

That is, continuation payoffs for lender and investors (including the lump sum transfer $dC(x)$) must be positive. We note $W_0(x)$ and $dC(x)$ do not depend on hidden and non-contractible screening effort q, but solely on the type x.

As before, we assume that the lender rejects the borrower and does not originate a loan if the signal is $S = b$ and identifies a bad type borrower in which case continuation surplus becomes F_B. Then, at time $t = 0^-$, the lender's private choice of effort is characterized by:

$$\max_{q \in [0, \bar{q}]} \left\{ \Big(\omega + (1-\omega) f(q)\Big) \Big[\hat{\omega}(q)(dC(g) + W_0(g)) + (1 - \hat{\omega}(q))(dC(b) + W_0(b))\Big] \right. \\ \left. + \Big[1 - \Big(\omega + (1-\omega) f(q)\Big)\Big] F_B - \frac{\kappa q^2}{2} \right\}. \quad \text{(B.95)}$$

To understand this objective, notice that — given the screening level q — $\omega + (1-\omega)f(q)$ is the

probability that the lender receives a positive signal $S = g$ and originates a loan. Conditional on origination, the borrower is good with probability $\hat{\omega}(q)$ — in which case the rating reveals $R = g$ — and bad with probability $1 - \hat{\omega}$ — in which case $R = b$. We would like to stress that the rating reveals borrower type *after* the loan has been originated and the borrower has been accepted. That is, the lender cannot reject the borrower anymore at the time of the rating, in line with common practice in the syndicated lending market that the rating is solicited after the lead arranger and other participating banks have committed credit to the borrower.

Next, we note that $F_0(x) = dC(x) + W_0(x)$ — that is, the continuation surplus after the rating reveals type x — does not depend on q. We can thus rewrite (B.95) as

$$\max_{q\in[0,\bar{q}]} F_{0^-} = \max_{q\in[0,\bar{q}]} \left(F_0 - \frac{\kappa q^2}{2} \right)$$

with $F_{0^-} = F_0 - \frac{\kappa q^2}{2}$ and

$$F_0 = \Big(\omega + (1-\omega) f(q)\Big)\Big[\hat{\omega}(q) F_0(g) + (1 - \hat{\omega}(q)) F_0(b)\Big] + \Big[1 - \Big(\omega + (1-\omega) f(q)\Big)\Big] F_B.$$

Thus, it becomes now apparent that the agent chooses screening effort q to maximize total surplus. That is, there is no more moral hazard over screening when the credit rating is perfectly informative. As a result, assuming a perfectly informative credit rating is akin to assuming that there is no moral hazard over screening (i.e., screening is contractible), as we do in the baseline in Section 4.2.

Continuation contract after rating. Finally, we characterize the continuation contract, following the perfectly informative credit rating that reveals borrower type x. As shown above, with this perfectly informative rating, there is no more screening moral hazard. Therefore, similar to the baseline solution in Section 2.2.1 without screening moral hazard, the solution — conditional on borrower type x being known — is stationary, featuring time-constant monitoring $a(x)$, continuation value $W(x)$, and continuation surplus $F_t = F(x)$ (with a slight abuse of notation $F(x)$ is not a function of screening incentives but of borrower type).

Using the integral representation of total surplus in (B.92) (which remains valid) and the dynamic programming principle, total surplus $F_0(x)$ is characterized by the following HJB equation:

$$F(x) = \max_{W(x), a(x), W^{Post}(x)} \left\{ \frac{1 - \frac{\phi a(x)^2}{2} - (\gamma - r) W(x) + \rho p^x F^{Post}(x)}{r + \rho} \right\},$$

with $p^B = 1 - \chi_B a(B)$ and $p^G = \overline{p} - \chi_G a(B)$, the monitoring incentive conditions, $a(B) = \frac{\rho \chi_B W^{Post}(B)}{\phi}$ and $a(G) = \frac{\rho \chi_G W^{Post}(G)}{\phi}$, as well as the monotonicity constraints $W^{Post}(x) \geq W(x)$ and $W^{Post}(x) \leq F^{Post}(x) - F(x) + W(x)$ (analogous to (B.93)) Here, $W^{Post}(x)$ and $F^{Post}(x)$ are continuation value for the lender as well as continuation surplus after the credit shock is survived, conditional on borrower type x.

Holding other controls equal, it is optimal to minimize $W(x)$ subject to the constraint (B.93), i.e., $W(x) \geq F(x) + W^{Post}(x) - F^{Post}(x)$, and $W(x) \geq 0$. As a result, $W(x) = \max\{F(x) +$

$W^{Post}(x) - F^{Post}(x), 0\} = (F(x) + W^{Post}(x) - F^{Post}(x))^+$. Thus,

$$F(x) = \max_{a(x), W^{Post}(x)} \left\{ \frac{1 - \frac{\phi a(x)^2}{2} - (\gamma - r)(F(x) + W^{Post}(x) - F^{Post}(x))^+ + \rho p^x F^{Post}(x)}{r + \rho} \right\},$$

with $p^B = 1 - \chi_B a(B)$ and $p^G = \overline{p} - \chi_G a(B)$, and the monitoring incentive conditions, $a(B) = \frac{\rho \chi_B W^{Post}(B)}{\phi}$ and $a(G) = \frac{\rho \chi_G W^{Post}(G)}{\phi}$. To conclude the characterization of the optimal contract with credit ratings, one can solve this optimization for $W(x)$ and $a(x)$.

To conclude, we first have micro-founded the reduced-form assumption of Section 4.2 that credit rating removes moral hazard over screening and, second and as in Section 4.2, that the optimal contract features time-stationary monitoring, continuation value (i.e., the agent's stake), and continuation surplus. The interpretation of this time-constant lender stake is that there are no more loan sales after origination.

B.6.10 Multiple Credit Shocks

This section shows that it is possible to extend the model to incorporate multiple credit shocks, at the expense of losing tractability and greatly complicating the analysis without adding new economic insights. Below, we sketch the solution with infinitely many credit shocks, each occurring at some intensity $\rho > 0$.

Upon credit shock at t, the loan defaults with probability p_t^x, with the default probability depending on the borrower type x. Suppose now a credit shock hits at time t and "just before" the credit shock at time t^-, the lender believes the borrower is type $x = G$ with probability $\hat{\omega}_{t^-}$. Then, if the loan survives this credit shock at t, the lender updates its belief according to Bayes' rule:

$$\hat{\omega}_t = Pr(x = G | \text{survive at } t, \hat{\omega}_{t^-}) = \frac{Pr(\text{survive at } t \ \wedge \ x = G | \hat{\omega}_{t^-})}{Pr(\text{survive at } t | \hat{\omega}_{t^-})} = \frac{\hat{\omega}_{t^-} p_t^G}{\hat{\omega}_{t^-} p_t^G + (1 - \hat{\omega}_{t^-}) p_t^B}.$$

We can more compactly write $\hat{\omega}_t$ as a function $\Omega(\cdot)$ of a_t, q, and $\hat{\omega}_{t^-}$:

$$\hat{\omega}_t = \Omega(a_t, q, \hat{\omega}_{t^-}).$$

The belief $\hat{\omega}_t$ now emerges as an additional state variable that is governed by a piece-wise constant process that only changes when a new shock hits and the firm survives.

From the lender's point of view, the expected probability of default is

$$p_t = \hat{\omega}_t p_t^G + (1 - \hat{\omega}_t) p_t^B,$$

which we can write as a function $p_t = p(a_t, q, \hat{\omega}_t)$. Using (B.88), which still applies in this context, the law of motion for W_t becomes

$$\dot{W}_t = (\gamma + \rho) W_t + \frac{\phi a_t^2}{2} - c_t - \rho(1 - p_t) W_t^{Post},$$

where W_t^{Post} is the continuation payoff after the next credit shock survival. Again, we assume that absent credit shocks and after time $t = 0$, payouts to the lender are smooth, that is, $dC_t = c_t dt$.

Monitoring incentives are, as before, $a_t = W_t^{Post}/\phi$.

Next, we differentiate $\dot{W}_t$ with respect to q to obtain the law of motion of $V_t = \frac{\partial}{\partial q} W_t$ which as a state variable characterizes screening incentives. The law of motion for V_t then becomes

$$\dot{V}_t = (\gamma + \rho)V_t - \rho(1 - p_t)V_t^{Post} + \rho p_{t,q} W_t^{Post}$$

where we define $p_{t,q} = p_q(a_t, q, \hat{\omega}_t) = \frac{\partial}{\partial q} p_t = \frac{\partial}{\partial q} p(a_t, q, \hat{\omega}_t)$ and $V_t^{Post} = \frac{\partial W_t^{Post}}{\partial q}$.

The two state variables for the dynamic optimization are $V_t = V$ and $\hat{\omega}_t = \hat{\omega}$; we omit time subscripts unless necessary. As in the model above and the baseline model, W and W^{Post} become control variables. As such, total surplus is a function of V and $\hat{\omega}$ only, i.e., $F_t = F(V_t, \hat{\omega}_t)$. Using (B.92) and the dynamic programming principle, the HJB equation in state $(V, \hat{\omega})$ is:

$$\begin{aligned}(r + \rho)F(V, \hat{\omega}) = \max_{a \in [0,\bar{a}], W, W^{Post}, V^{Post}} \Bigg\{ & 1 - \frac{\phi a^2}{2} - (\gamma - r)W + \rho(1 - p(a, q, \hat{\omega}))F\big(V^{Post}, \Omega(a, q, \hat{\omega})\big) \\ & + F_V(V, \hat{\omega})\Big((\gamma + \rho)V - \rho p_q(a, q, \hat{\omega})W^{Post} - \rho(1 - p(a, q, \hat{\omega}))V^{Post}\Big)\Bigg\},\end{aligned}$$

where the optimization is subject to (B.90), (B.93), $W^{Post} \geq W_t$, and $W \geq 0$ and $F_y(V, \hat{\omega}) = \frac{\partial}{\partial y} F(V, \hat{\omega})$. Because there are infinitely many credit shocks and after the credit shock survival in state $(V, \hat{\omega})$ the new state becomes $V^{Post}, \Omega(a, q, \hat{\omega})$, we obtain that $F^{Post} = F\big(V^{Post}, \Omega(a, q\hat{\omega})\big)$. The initial choice of screening as well as the incentive condition for screening is as in Section B.6.7. Overall, the dynamic optimization becomes involved and complicated. As such, we do not proceed from here, in that a complete analysis of this extension is beyond the scope of the paper.

B.7 Upper Bounds on Effort

Throughout the paper, we have assumed (and implicitly focused on parameterizations ensuring) that the exogenous upper bounds on effort $\bar{a}$ and $\bar{q}$ do not bind. Moreover, we have imposed parameter conditions such that the first order approach is valid and the agent's objective function is strictly concave in its choice of efforts (see Lemma 1).

With these assumptions, incentive conditions regarding screening and monitoring efforts are simply the agent's first order condition, creating a one-to-one link between incentives and effort levels.

It is straightforward to extend the analysis to the case that the inequalities $a_t \leq \bar{a}$ or $q \leq \bar{q}$ may bind, notably, without changing the model's qualitative implications. Allowing for this possibility, it is clear that the incentive condition for monitoring (see (6)) would change to

$$W_t \geq a_t \phi,$$

where the inequality is tight for $a_t < \bar{a}$. Likewise, the incentive condition for screening (see (9)) would change to

$$V_0 \geq \kappa q,$$

where the inequality is tight for $a_t < \bar{a}$.

Analogous to the baseline treatment in the main text, the maximization of total surplus is then characterized by the HJB equation

$$rF(V) = \max_{a\in[0,\bar{a}],W} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)\big((\gamma+\lambda)V - W\big) \right\}, \tag{B.96}$$

while the HJB equation is solved subject to the constraints $W \geq a\phi$ (with the inequality being tight for $a < \bar{a}$) and $W \in [0, F(V)]$.

When $a \in (0, \bar{a})$, we have $W = \phi a$ and the first order condition with respect to a and $W \in [0, F(V)]$ pin down

$$\frac{F(V) - F'(V)[V+\phi] - (\gamma - r)\phi}{\phi} \quad \wedge \quad \frac{F(V)}{\phi}.$$

Otherwise, when $a = \bar{a}$, we have $W \geq \phi\bar{a}$ and

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

The initial optimization with respect to q then becomes

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

with the inequality being strict if $q < \bar{q}$. As in the baseline model, when $F'(V) \leq 0$, then $V_0 = \kappa q$ is optimal.

Thus, when $a_t \leq \bar{a}$ and $q \leq \bar{q}$ may bind, the dynamic optimization has to distinguish between different cases, i.e., $a < \bar{a}$ and $a = \bar{a}$, making the analysis more complicated but not changing the economic forces at work. To reduce the number of different cases to consider and to streamline the exposition, we therefore assume throughout the paper that the upper bounds on effort do not bind.

B.8 Model Variant with only Moral Hazard over Screening

B.8.1 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 (6) 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 follows[45]

$$dW_t = (\gamma + \lambda_t)W_t dt + \frac{\phi a_t^2}{2} dt - dC_t, \tag{B.97}$$

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 (B.97) with respect to screening effort q to obtain (after simplifications) for $V_t = \frac{\partial}{\partial q} W_t$:

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

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

$$rF(V) = \max_{W \in [0, F(V)], a \in [0, \bar{a}]} \left\{ 1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F(V) + F'(V)((\gamma + \lambda)V - W) \right\}, \tag{B.98}$$

which is analogous to the baseline HJB equation (23). 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 (B.98). In what follows, we assume that a unique solution to (B.98) (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 (B.98) 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 (B.98) yields that

$$W(V) \begin{cases} = 0 & \text{if } F'(V) > -(\gamma - r) \\ \in [0, F(V)] & \text{if } F'(V) = -(\gamma - r) \\ = F(V) & \text{if } F'(V) < -(\gamma - r). \end{cases} \tag{B.99}$$

Note now that when screening is observable and contractible (in addition to monitoring being observable and contractible), then $V^B(q) = W^B(q) = 0$. As in the baseline, it follows that $\lim_{t \to \infty} V_t = V^B(q) = 0$, i.e., given q, the optimal contract approaches in the limit $t \to \infty$ the

[45] Since 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.

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 (B.99), 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 (B.98) to obtain

$$\gamma F(V) = \max_{a \in [0,\bar{a}]} \left\{ 1 - \frac{\phi a^2}{2} - \lambda F(V) + F'(V)((\gamma + \lambda)V - F(V)) \right\}. \tag{B.100}$$

Let us assume that $F''(V)$ exists and is well-defined. Using the envelope theorem, we totally differentiate the HJB equation (B.100) (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 (B.99) 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 (B.98). Notice that this boundary condition is equivalent to

$$F(0) := \lim_{V \to 0} F(V) = \max_{a \in [0,\bar{a}]} \left(\frac{1 - \frac{\phi a^2}{2}}{r + \Lambda - a - q} \right), \tag{B.101}$$

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 $\tau^0 = \inf\{t \geq 0 : V_t = 0\}$. Once V has reached zero, the contract remain stationary and implements constant effort $a(0)$, while total surplus reads $F(0)$ (defined above) and the agent's stake remains constant.

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 $V_0 = \kappa q$.

B.8.2 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 of the loan 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 $dC(c) = F(0)$ at the time V reaches zero (i.e., at time τ^0), so as to induce $F(0) = \lim_{V \downarrow 0} W(V) > W(0) = 0$. After time τ^0 (i.e., for $t > \tau^0$ once V_t has reached zero), W_t remains constant at zero so that $dC_t = \frac{\phi a_t^2}{2} dt$.

When $V > 0$, then $F(V) = W(V)$, and according to (A.14) for $W(V) = F(V)$:

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

yielding

$$dC = 1dt,$$

which equals coupon payments over an instant dt.

We implement the optimal contract by having the agent retain a share β_t of the loan for $t \geq 0$. After time τ^0, i.e., for times $t > \tau^0$, we have $dC_t = \frac{\phi a(0)^2}{2} dt$ and the agent retains a constant share of the loan

$$\beta(0) := \beta_t = \frac{\phi a_t^2}{2} = \frac{\phi a(0)^2}{2}.$$

That is, after time τ^0, the agent retains stake $\beta(0)$ which implements payouts such that the agent is compensated for its cost of monitoring effort.

Next, for times $t \in (0, \tau^0)$, we have $dC_t = 1dt$ and the agent's share β_t satisfies

$$\beta_t + (-\dot{\beta}_t)L_t = 1, \tag{B.102}$$

where $L_t = \int_t^\infty e^{-r(s-t) - \int_t^s \lambda_u du} 1 ds$ is the fair market price of the loan and subject to β_0 given.Notice that with starting value $\beta_0 = 1$, the ODE (B.102) implies $\dot{\beta}_t = 0$ so that $\beta_t = 1$ for $t \in [0, \tau^0)$. As a result, the contract is implemented by requiring the agent to fully retain the pool of loans until time $\tau^0 = \inf\{t \geq 0 : V_t = 0\} < \infty$.

When V reaches zero at time τ^0, the agent sells fraction $1 - \beta(0)$ of the loan to the principal (outside investors), and it receives the fair (per-unit) price of $L(0) = \frac{1}{r + \Lambda - a(0) - q}$ dollars, i.e., in total $(1 - \beta(0))L(0)$. This lumpy loan sale implements the agent's contracted lump-sum payout

$$dC(0) = F(0) = \frac{1 - 0.5\phi a(0)^2}{r + \Lambda - a(0) - q} = \frac{1 - \beta(0)}{r + \Lambda - a(0) - q} = (1 - \beta(0))L(0)$$

at time τ^0 (i.e., in state $V = 0$). Notice that the agent sells its entire stake at τ^0 when $a(0) = 0 \iff \beta(0) = 0$, i.e., when $\phi = 0$, $\phi = \infty$, or $\bar{a} = 0$. By construction, these retention dynamics implement the contracted payments to the agent and as such the optimal contract.