The Review of Financial Studies

The Dynamics of Loan Sales and Lender Incentives

Sebastian Gryglewicz
Erasmus University Rotterdam, Netherlands

Simon Mayer
Tepper School of Business, Carnegie Mellon University, USA

Erwan Morellec
EPF Lausanne, Swiss Finance Institute, Switzerland, and CEPR

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. (*JEL* G21, G32)

Received: October 10, 2022; Editorial decision: January 3, 2024
Editor: Itay Goldstein.
Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

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, and 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 was 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." Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Erwan Morellec, erwan.morellec@epfl.ch.

The Review of Financial Studies 00 (2024) 1–58

https://doi.org/10.1093/rfs/hhae021 Advance Access publication May 24, 2024

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 et al. 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 et al. 2021). As further shown by Blickle et al. (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 share and its relation to moral hazard and loan performance. This allows us to (a) shed light on recent empirical findings and (b) 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 nonbank financial intermediaries. The loan generates coupon payments at a constant rate until default or maturity. The lender may undertake a costly screening effort at

[1] Syndicated loans are loans jointly issued to a borrower 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; Saunders et al. 2021).

[2] As documented in, for instance, 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 other institutional investors and banks. Furthermore, each loan included in CLOs is rated.

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, for example, 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 sell-off speed decreases over time, so most loan sales occur 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 sell-off 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 sell-off dynamics, consistent with the evidence in Chen et al. (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

[3] Exploiting plausibly exogenous shocks to the severity of moral hazard in loan origination, Chen et al. (2023) and Jiang, Kundu, and Xu (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 nonbank intermediaries. Haque, Mayer, and Wang (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 to retain a lower loan share and to sell more loan shares to nonbank intermediaries.

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), (a) most loan sales occur relatively shortly after origination, and (b) the lender may sell the entire loan in finite time, that is, before maturity. As we also show analytically, the latter scenario prevails when the benefits of monitoring (relative to its costs) are limited or the lender's cost of capital is high, generating large gains from trade. Interestingly, our findings reveal that when the lender's cost of capital is sufficiently high, the lender may sell the entire loan before maturity, while exerting significant screening and monitoring efforts that lead to a sizeable reduction in credit risk. That is, the fact that the lender sells the entire loan before maturity does not necessarily imply that the lender adds little value through its screening and monitoring. Indeed, the opposite can be true.

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 (prescreening) credit risk implies earlier default and thus both a shorter time period 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, Ivanov, and Meisenzahl (2021)—document a negative relation between screening—respectively monitoring—and credit risk, our results point to a two-way causality. Screening and monitoring reduce credit risk, and 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 sell-off 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 nonmonotonic 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 sell-off 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, Ivanov, and Meisenzahl (2021). We additionally show that while sell-off speed is monotonic in the level of monitoring effort, it is nonmonotonic in the level of screening effort. The nonmonotonic relationships between sell-off speed and screening as well as between initial retention and monitoring imply that neither initial retention nor a measure of sell-off 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. 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. We also examine retention dynamics when the originator cannot commit to a specific retention level. We show that the sell-off dynamics and how they are affected by model parameters are *qualitatively* similar in the zero- and full-commitment solution, so that our analysis allows us to draw robust inferences on how loan and lender characteristics shape sell-off dynamics under moral hazard.

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; 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, Dlugosz, and Ivashina 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 sell-off 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 et al. (2007). In this literature, Piskorski and Westerfield (2016), Malenko (2019), Orlov (2022), and Gryglewicz and Mayer (2022) analyze incentive provision with optimal dynamic contracts and monitoring. Halac and Prat (2016), Varas, Marinovic, and Skrzypacz (2020), and Hu and Varas (2021) characterize optimal monitoring in dynamic settings but do not focus on optimal contracts. In a related paper, Hartman-Glaser, Piskorski, and 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, Piskorski, and Tchistyi (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 corporate loans are both screened and monitored in practice, 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 share of the loan. Notably, unlike Hartman-Glaser, Piskorski, and Tchistyi (2012) or Hoffmann, Inderst, and Opp (2021, 2022), our model generates retention and sell-off 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 findings and to generate new predictions regarding optimal retention, loan sales dynamics, 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, for example, 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, for example, to more information collected and processed during the due diligence process and, thus, to lower levels of default risk.[4]

Monitoring effort $(a_t)_{t\geq 0}$ captures the lender's post-origination due diligence and monitoring, which can take various forms (for direct evidence on bank monitoring, see, e.g., Gustafson, Ivanov, and Meisenzahl 2021; Heitz, Martin, and Ufier 2022). 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 the evidence in Heitz, Martin, and Ufier (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 preorigination 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 Internet 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.[5] If the lender decides to shirk on either task, the loan will have a higher default rate. Although

[4] 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, Piskorski, and Tchistyi (2012), Malamud, Rui, and Whinston (2013), and Hoffmann, Inderst, and Opp (2021). Internet 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.

[5] 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 Internet Appendix B.1 for a theoretical analysis of this model variant, Internet Appendix B.6 for its micro-foundation, and Section 2.2.4 for a numerical analysis.

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 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.[6] 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).

Because of 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.[7] 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 $t > 0$.

1.3 Contracting problem

In what follows, we write $t = 0^-$ as the time just before the screening effort is chosen, and $t = 0$ denotes 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

[6] 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 as the risk-neutral or "risk-adjusted" default probability.

[7] Section 4.4.1 shows how commitment can arise through repeated originator-investor relations.

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.[8] 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 the lender and investors are then formally defined as $W_t \geq 0$ and $P_t \geq 0$ for any $t \geq 0$.[9]

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 $t \geq 0$.

[8] 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 to zero, it follows that $dC_t = 0$ for $t > \tau$.

[9] 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.

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 characterize the lender's incentives for screening and monitoring, and hence the resultant 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. Because of limited liability, the agent only loses its claim to future payments, that is, 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

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

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 because of its relative impatience ($\gamma > r$).

While monitoring a_t affects 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, $\frac{\kappa q^2}{2}$:

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

V_t denotes 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:[10]

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

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 to 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, that is, $V_t = D_t W_t$. 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

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

[10] When differentiating W_t, we can ignore the effect on a_t because of 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.

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.

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 noncontractible 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[11]

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

$$= \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}\left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s\right)ds, \tag{20}$$

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

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

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

[11] 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).

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).[12] Equation (22) is solved subject to the incentive condition (6), the limited liability constraints, and the conjecture that payouts to the lender are smooth, in that $dC=cdt$. Note that it is always possible to stipulate that the lender receives an incremental payout of Δ dollars,[13] 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

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

[12] 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 $F_t=F(V_t,W_t)$.

[13] 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 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), that is, $W=\phi a$, and the principal's and the agent's limited liability conditions, that is, $W\in[0,F(V)]$.

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 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.

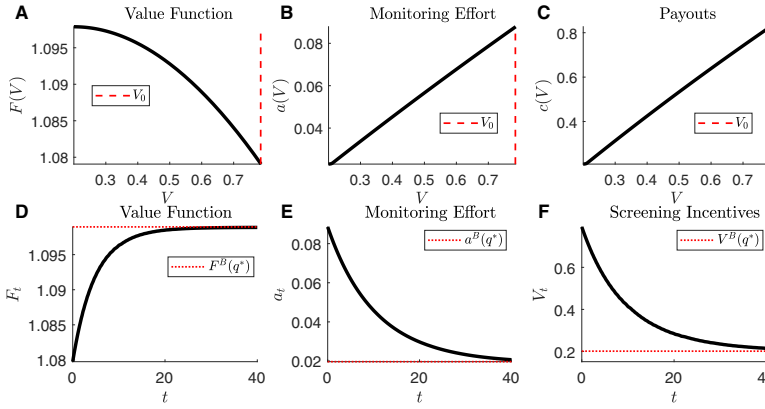

Figure 1
Total surplus $F(V)$, monitoring $a(V)$, and the agent's flow payouts $c(V)$
In the upper panels, the vertical dashed red line represents V_0. In the lower panels, the horizontal dotted red line represents the benchmark levels that are attained in the limit $t \to \infty$.

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 modeled 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.[14]

2.2.3 Contract dynamics. Figure 1 provides a numerical example of the optimal contract. For the numerical analysis, we normalize $r = 0$ and $\Lambda = 1$ so that, without monitoring and screening, the expected time to default is $1/\Lambda = 1$ year and the loan has a pre-effort (or intrinsic) value $1/(\Lambda + r) = 1$.[15] 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, that is, 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.

[14] See also Section B.6.8 in the Internet Appendix.

[15] 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).

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 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 Internet Appendix B.1. A micro-foundation of this default intensity can be found in Internet 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

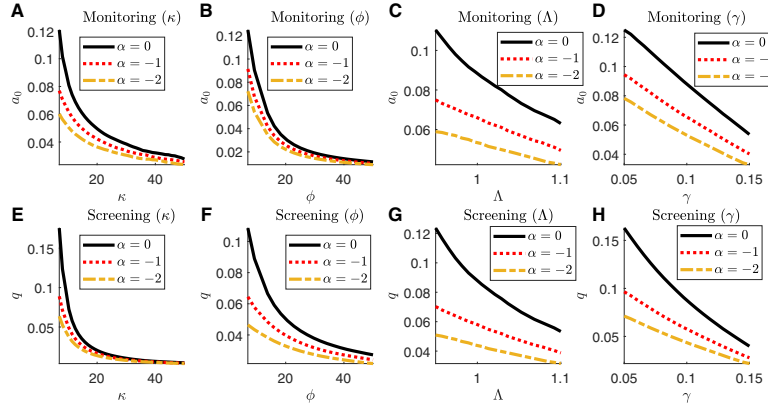

Figure 2
Comparative statics
This figure plots the monitoring effort a_0 for $\alpha = 0$ (solid black line) and screening effort q^* against the parameters ϕ, κ, and Λ for $\alpha = -1$ (dotted red line) and $\alpha = -2$ (dashed yellow line). We use our baseline parameters.

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.[16]

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 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.[17]

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

[16] 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.

[17] 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.

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

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

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, that is, 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 $\phi \to \infty$,[18] monitoring is prohibitively costly, so both contracted and actual monitoring equal zero and, mechanically, there is no moral hazard over monitoring.[19] 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 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, Internet 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, that is, 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) and Irani et al. (2021).[20] 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.

---

[18] 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 $\Lambda - \bar{a}$.

[19] 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 $\phi \to \infty$.

[20] Loan sales and securitization are common in other markets too and affect lender incentives. For instance, Purnanandam (2011) and Keys et al. (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.

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 coinvestors (other banks in the syndicate) jointly commit the loan to the borrower, with loan terms being determined based on the screening outcome.[21]

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, Malherbe, and Meisenzahl 2020)—outside investors may buy loan shares right away or commit to buying loan shares in the secondary market.[22] That is, during the primary market stage, the lead arranger gradually reduces its exposure to the loan by engaging in loan sales or, alternatively, precommitted loan sales (akin to a forward sale of the loan). During this primary market stage, the lead arranger is exposed to pipeline risk, that is, the risk that it cannot sell the loan if investor demand dwindles, for example, because of bad news about the borrower (not necessarily limited to actual default).[23] 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 precommitted 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 markets), 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

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

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

[23] 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 precommitted loan sales.

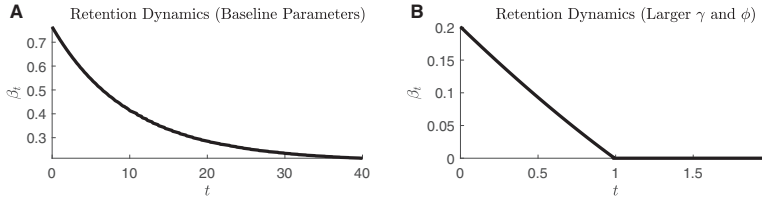

Figure 3
Implementation of the optimal contract and per unit value of the loan
Panel A uses our baseline parameters, and panel B sets $\gamma = 0.18$ and $\phi = 11$.

argue next, the retention and loan sales dynamics qualitatively resemble the patterns observed in the data.

First, observe that the sell-off speed, as captured by $-\dot{\beta}_t$ in Figure 3, decreases with time t since origination. That is, in panel A, β_t is convex and decreasing in t approaching β_B in the limit (the sell-off speed tends to zero as $t \to \infty$), while in panel B the sell-off 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, Liu, and Stebunovs (2022). In fact, under certain parameter conditions, the lender sells off its entire stake in finite time, capturing the sell-off 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. Panel B of Figure 3 plots the retained share when lender's discount rate and the cost of monitoring are larger than in the baseline. The lender retains initially $\beta_0 \approx 20\%$ of the loan—in line with the initial retention level reported in Sufi (2007)—and sells its entire stake in finite time in about 1 unit of time (i.e., one year corresponding to the median sell-off time for term B loans reported in Blickle et al. 2022).

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 the loan in finite time. Thus, our results reveal that lenders retain loans that have a low holding cost γ or to which they can add value through monitoring (because of low ϕ), while selling loans for which the holding cost is high relative to the potential value added through monitoring.

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

1. When the cost of monitoring ϕ or the cost of capital γ are 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 ϕ or cost of capital γ are sufficiently low in that $\phi < \frac{1}{(r+\Lambda-q)(\gamma-r)} \leq \frac{1}{(r+\Lambda)(\gamma-r)}$, the lender never sells its entire stake,

that is, $\beta_t > 0$ and $W_t > 0$ for all $t \geq 0$. Thus, $\phi \geq \frac{1}{(r+\Lambda-q)(\gamma-r)} \geq \frac{1}{(r+\Lambda)(\gamma-r)}$ is a necessary condition for sell-off in finite time.

Importantly, the sufficient condition (30) for full sell-off does 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 that 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 sell-off (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, for example, Hartman-Glaser, Piskorski, and Tchistyi (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 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.[24]

Finally, we would like to highlight that evidence points toward moral hazard as an important driver of loan sale dynamics and vice versa. Gustafson, Ivanov, and Meisenzahl (2021) document that the extent of active monitoring crucially

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

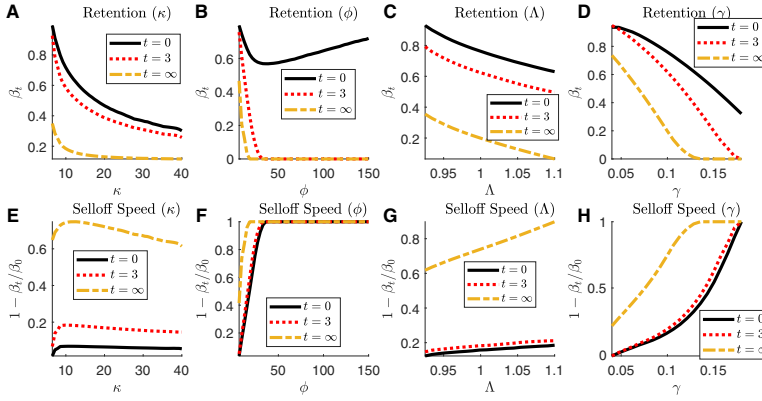

Figure 4
Retention and dynamics
Initial retention and sell-off speed as functions of the costs of screening and monitoring κ and ϕ, intrinsic credit quality Λ, and the lender's cost of capital γ.

depends on the lead arranger's retained share and loan sales. Chen et al. (2023) and Haque, Mayer, and Wang (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 sell-off 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 sell-off 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).

Figure 4 reveals that retention decreases and sell-off 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, Gerardi, and Hartman-Glaser (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 sell-off speed. Initial retention decreases with κ. However, the sell-off speed is hump-shaped in κ.[25] 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 sell-off speed is the highest for intermediate κ, note that when κ is sufficiently low, moral hazard over screening becomes negligible and the optimal contract only needs to incentivize monitoring. Thus, the contract comes close to that in the benchmark with only monitoring moral hazard and a constant level of retention, that is, a zero sell-off 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 sell-off speed. Consequently, screening effort, which is monotonically decreasing in κ, can be either increasing or decreasing in sell-off speed.

Panels B and F of Figure 4 show the relation between the cost of monitoring ϕ and the levels of retention and sell-off speed. Remarkably, in contrast to the effect of κ, initial retention is nonmonotonic 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 sell-off 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 sell-off speed. As noted above, monitoring effort is nonmonotonic in initial retention and screening effort is nonmonotonic in sell-off speed. Because loan performance depends on both screening and monitoring, these nonmonotonic relations help rationalize the finding of Blickle et al. (2022) that initial retention or sell-off 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

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

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 sell-off dynamics after time t, but not directly on β_0 or the loan sales up to time t. In line with our theory, Gustafson, Ivanov, and Meisenzahl (2021) find that monitoring in a given year is positively related to the lead share in the same year.

3.4 Loan sales and the effects of screening and monitoring

What do loan sales imply for the value of screening and monitoring? Does the fact that the lead arranger sells the entire loan before maturity mean that screening and monitoring have little or no effect on credit risk? Interestingly, we can use Corollary 1 to derive an upper bound on the effect of monitoring (and screening) on credit risk. A loan sale in finite time implies the (necessary) condition $\phi \geq \frac{1}{(r+\Lambda-q)(\gamma-r)}$ under the optimal level of screening q. We know that at any point time $W_t < \frac{1}{r+\Lambda-q}$, so that $a_t < \frac{1}{(r+\Lambda-q)\phi}$. Combining these conditions yields $\gamma - r > a^{Max} := \max_{t\geq 0} a_t$. Monitoring reduces total credit risk captured by the default intensity $\Lambda - a_t - q$ at time t maximally by a^{Max} in absolute terms, as absent monitoring, the default intensity would be $\Lambda - q$. In addition, as $\frac{1}{\Lambda-q-a^{Max}} - \frac{1}{\Lambda-q} = \frac{a^{Max}}{(\Lambda-q)^2} + o(a^{Max})^2$, we have

$$\frac{1}{\Lambda-q-a^{Max}} - \frac{1}{\Lambda-q} \approx \frac{a^{Max}}{(\Lambda-q)^2}.$$

Monitoring therefore maximally decreases the expected time to default by $\frac{a^{Max}}{(\Lambda-q)^2}$ in absolute terms and by $\frac{a^{Max}}{\Lambda-q}$ in relative terms. We have shown that $\gamma - r > a^{Max}$. Clearly, we also have $\frac{1}{\Lambda-q} \leq \overline{\tau}$ because additional monitoring increases the expected time to default $\overline{\tau}$. As a result, when a loan is sold off in finite time, our model implies

$$(\gamma - r)\overline{\tau} > \frac{a^{Max}}{\Lambda-q}.$$

The left-hand side is the product of the lender's effective cost of capital (relative to the risk-free rate) and the expected time to default (at origination). The right-hand-side proxies for the maximum reduction in credit risk stemming from monitoring (i.e., monitoring reduces credit risk by $100 \cdot \frac{a^{Max}}{\Lambda-q}$ percentage points). The values of γ, r, and $\overline{\tau}$, therefore, imply an upper boundary on the relative change in the expected time to default through monitoring. As such,

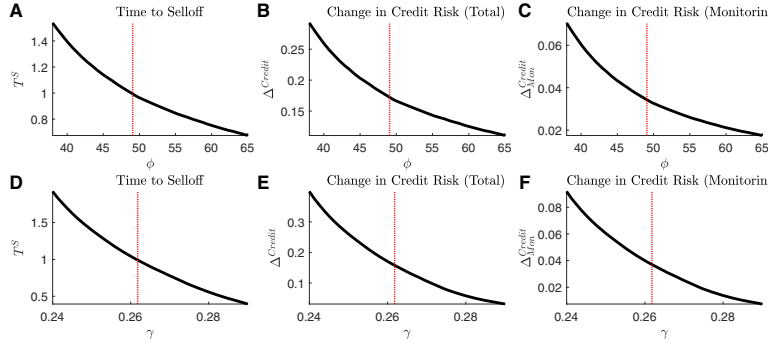

Figure 5
Loan sales and the effects of screening and monitoring
The parameters are $r=0.05$, $\gamma=0.25$, $\Lambda=0.3$, $\delta=0.25$, $\phi=\kappa=40$, and $1/\delta=3.33$. The vertical dotted red line depicts the point at which time to sell-off T^S equals one. The parameter δ is formally introduced in Section 4.3, where we analyze finite maturity.

the fact that a loan is sold in a finite time necessarily suggests, in line with our previous findings, that monitoring has a limited impact on reducing credit risk or that the lender's cost of capital and the gains from trade are relatively large. However, as the cost of capital of the lender can indeed be large, a loan sale in finite time does not imply that monitoring, let alone screening, has a low quantitative impact on credit risk.

To illustrate this point, Figure 5 presents a numerical example with a high lender cost of capital, in which the loan sale occurs in a finite time before maturity, but screening and monitoring have significant effects on credit risk. For this figure, we set $r=0.05$, $\Lambda=0.3$, $\gamma=0.25$, and $\phi=\kappa=40$; we interpret one unit of time as one year. In addition, we introduce a finite maturity of 3.33 years in line with the average maturity reported in Sufi (2007). (Section 4.3 generalizes our model to discuss the effects of loan maturity). This parameter choice leads to the sell-off of the entire loan before maturity, which we view as a corner case of our model and requires assuming sufficiently large gains from trade $\gamma-r=0.2$ as shown formally in Corollary 1. Note that the lender's large(r) discount rate γ may reflect not only capital or financial constraints but also preferences (e.g., risk aversion) or an outside option that delivers high returns.

The upper three panels of Figure 5 present comparative statics in ϕ for $\phi\in[38,65]$ while the lower three panels present comparative statics in γ for $\gamma\in[0.24,0.29]$. Panels A and D plot the time to sell-off (in years) defined as $T^S=\inf\{t\geq 0:\beta_t=0\}$. Observe that the expected time to sell-off ranges from about 0.4 to 1.8, and equals about one year for intermediate parameter values, in line with the median time to sell-off for term B loans reported in Blickle, Parlatore, and Saunders (2023) (with a time to sell-off for the average, not necessarily term B, loan being likely significantly larger). The vertical dotted red line shows the point at which time to sell-off equals one year. Panels B and E

plot the total relative increase in expected time to default $\overline{\tau}$ due to screening and monitoring, which is $\Delta^{Credit} := \frac{\overline{\tau}-1/\Lambda}{1/\Lambda} = \Lambda\overline{\tau}-1$, noting that absent screening and monitoring the expected time to default is $1/\Lambda$. The expected time to default is defined as $\overline{\tau} = \int_0^\infty e^{-\int_0^t \lambda_u du} dt$ and serves as a measure of credit risk, with a lower (higher) expected time to default corresponding to higher (lower) credit risk. Panels C and F plot the increase in the expected time to default due to monitoring, that is, $\Delta^{Credit}_{Mon} := \frac{\overline{\tau}-1/(\Lambda-q)}{1/(\Lambda-q)} = (\Lambda-q)\overline{\tau}-1$, noting that absent monitoring and given q the expected time to default is $1/(\Lambda-q)$.[26]

Observe that, as expected, the lender's propensity to sell its entire stake increases as ϕ or γ increases and thus T^S decreases (see panels A and D), in line with our previous findings. At the same time, the middle and right panels illustrate that the reduction in credit risk, because of either monitoring alone or screening and monitoring jointly, decreases with ϕ and γ. Thus, our analysis reveals that, indeed, screening and monitoring have a lower impact on credit risk when the lender sells its entire stake relatively quickly after origination.

However, as can be seen from the middle and right panels of Figure 5, a sell-off in finite time does not imply that the effects of screening and monitoring on credit risk are quantitatively small. In panel B, the reduction in credit risk due to screening and monitoring exceeds 10% in all scenarios considered, even when $\phi = 65$ and sell-off occurs in about 0.7 years. When the time to sell-off equals one year at the dotted red line, screening and monitoring reduce credit risk by more than 15%, while monitoring alone reduces credit risk by about 4% (see panel C). Likewise, in the lower three panels, when sell-off occurs within one year at the dotted red line, screening and monitoring reduce credit by about 15% and monitoring reduces credit risk alone by about 4%. That is, screening and monitoring can have significant effects on credit risk despite the sell-off before maturity.

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

[26] Admittedly, Δ^{Credit}_{Mon} is a conservative lower bound on the effect of monitoring on credit risk, as its determination takes the level of screening q as given and does not take into account the *complementarity* between screening and monitoring. With lower or no monitoring, the level of screening would be lower, too, due to this complementarity.

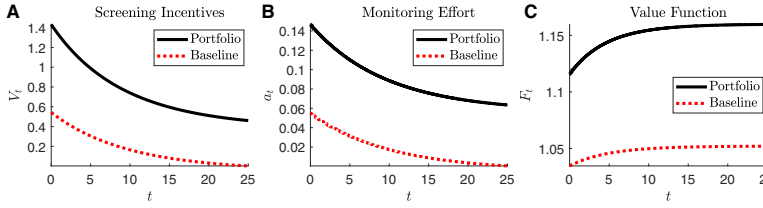

Figure 6
Time dynamics and loan portfolios
This figure plots screening incentives V_t, monitoring effort (per individual loan) a_t, and value function (per individual loan) against time t. We use our baseline parameters but consider a higher cost of screening and monitoring (i.e., $\phi = \kappa = 11$ and $\bar{a} = 0.2$) to ensure optimal efforts are interior also when considering portfolios.

times are independent, conditional on the lender-investor contract $\mathcal{C}$. The detailed description and solution of this model variant can be found in Internet 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, that is, 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 Internet 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 Internet 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.

Figure 6 illustrates the dynamics of incentives and payoffs both in our model variant with loan portfolios (solid black line) and the baseline (dotted red line). On a high level, the dynamics of incentives are similar in both model variants, leading to time-decreasing incentives and lender stake akin to loan sales. Panel A plots the lender's screening incentives and shows that, in both instances, screening incentives decrease over time.[27] In addition, monitoring efforts a_t, which are proportional to the lender's stake W_t, decrease over time

[27] Note that in the model variant with portfolios, the lender exerts the same monitoring and screening efforts for both loans, as the loans are symmetric. Further, we plot the outcomes prior to any default event.

too (see panel B). And, just as in Figure 1, the value function per loan (i.e., we divide the total value function by two in the loan portfolio case, as there are two loans) increases over time t in both scenarios but is significantly larger when a portfolio is originated. In both scenarios, screening and monitoring incentives decrease over time because the lender gradually reduces its stake in the loan, akin to gradual loan sales and time-decreasing retention. This suggests that also when the lender originates a portfolio of loans, it gradually reduces its stake and incentives. Although the lender has stronger overall incentives when it originates a portfolio of loans, the underlying moral hazard problem still persists and shapes the contract and loan sale dynamics, leading to time-decreasing lender stake and incentives.

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 Internet Appendix B.6. The intuition underlying this assumption is that the credit rating at origination reveals loan quality and generates screening incentives, as lax screening would lead to a low rating. Because the credit rating cannot be conditioned 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.

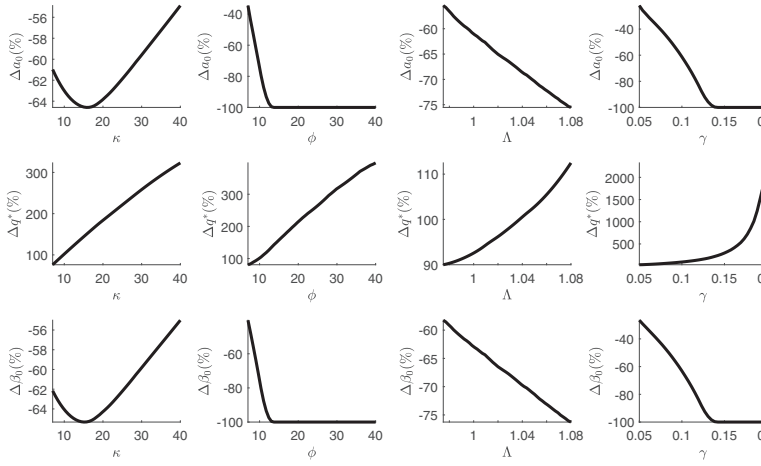

Figure 7
The effects of credit ratings
Δy denotes the percentage change in the initial value of the outcome variable y caused by a credit rating, where $y \in \{a_0, q^*, \beta_0\}$. Outcome variables are plotted as functions of the cost of monitoring κ, the cost of screening ϕ, the raw default intensity Λ, and the lender's discount rate γ.

Figure 7 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 7 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 Internet 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 Internet Appendix B.3, the agent's screening incentives at time $t = 0$ read

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

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 8 plots the initial monitoring effort (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 sell-off (panel D) after origination.

The effects of debt maturity on screening and monitoring feed back into default risk. Notably, panel E of Figure 8 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

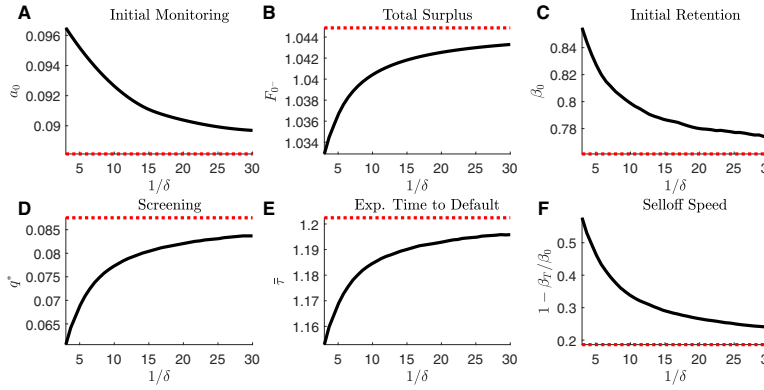

Figure 8
The effects of debt maturity
We use our baseline parameters and set $T=3$ for sell-off speed. The dotted red line depicts the outcomes with infinite debt maturity.

maturity have higher default risk.[28] Panel B of Figure 8 shows that total surplus increases with debt maturity due to lower agency costs.

4.4 Commitment

4.4.1 Commitment through repeated loan origination. Repeated lender-borrower and lender-investor interactions are common in credit markets, in particular in syndicated lending.Internet Appendix B.4 solves a model variant in which the lender originates a new and identical loan to a new borrower when the current loan matures so that the lender has to rescreen and exert costly screening effort when a new loan is originated. Internet Appendix B.4 derives 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, Nanda, and Yerramilli 2011).[29] Thus, with repeated interactions, the lender retains a lower share or sells the loan faster.

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

[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

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

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

[29] 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.

retain a time-decreasing share of the loan, thus inducing optimal loan sales as part of the optimal full-commitment contract. 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 Internet 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.

4.4.2 Loan sale dynamics absent commitment. What happens then in the absence of a commitment mechanism? In Internet Appendix B.9, we present a model variant in which the originator cannot commit to a specific retention level. In this variant, the originator's share of the loan at time t is denoted by $\beta_t \in [0,1]$, with $\beta_0 = 1$. The issuer can trade its stake at its own discretion by choosing $d\beta_t$ without commitment at any point in time t (before default). In addition, the issuer chooses, at any point in time, optimal monitoring $a_t \in [0, \bar{a}]$ and, at time $t = 0^-$, screening $q \in [0, \bar{q}]$.

As we show in Internet Appendix B.9, the lender's optimal trading strategy may involve both smooth and lumpy trading. Notably, we show that there exists an endogenous threshold $\overline{\beta}$ so that the lender sells its share smoothly for $\beta > \overline{\beta}$ at an endogenous rate $\dot{\beta} < 0$, that is, $d\beta = \dot{\beta} dt$. Once β reaches $\overline{\beta}$ from above, the lender randomizes at an endogenous rate between selling its entire stake at once or not trading at all, so β remains constant at $\overline{\beta}$ until it jumps to zero. That is, for $\beta = \overline{\beta}$, we get $d\beta = -\beta dN$ where $dN \in \{0, 1\}$ is a jump process with endogenous intensity $\mathbb{E}[dN] = \xi dt$. Last, when $\beta < \underline{\beta}$, the lender optimally sells its entire stake at once at price $\underline{L}$, that is, $d\beta = -\beta$. As we show, $\underline{\beta}$ can lie above 1, in which case the lender sells the entire loan at origination, leading to zero screening and monitoring.

A key object characterizing the dynamics of loan sales in this model variant is the threshold $\underline{\beta}$, which captures the lender's propensity to sell the (entire) loan. We show that $\underline{\beta} = 2\phi(\gamma - r)(\Lambda + r - q)$ so that this threshold increases with the cost of monitoring ϕ, the lender's cost of capital $\gamma - r$, the intrinsic credit risk Λ, and decreases with the level of screening q. That is, as in the optimal full-commitment contract, the lender is more likely to sell its entire stake when monitoring adds relatively little value and ϕ is large, when holding the loan on the book is costly, or when the loan is risky and Λ is high. In addition, a higher cost of screening κ reduces the screening level $q = q^*$ and thus boosts loan sales. Similar conclusions arise from an analysis of the state-dependent trading rate.

Different from the optimal contracting solution, monitoring and screening might not be feasible absent commitment, as the lender may sell the entire

loan at origination, which will happen when $\underline{\beta} > 1$. Then, clearly, the lack of commitment increases credit risk. Instead, when $\underline{\beta} < 1$, the sell-off dynamics and how they are affected by model parameters are *qualitatively* similar in both the zero- and full-commitment solution so that our analysis allows us to draw robust inferences on how loan and lender characteristics shape sell-off dynamics under moral hazard.

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). Internet 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 Internet 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 (prescreening) 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 sell-off 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 nonmonotonic 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 nonmonotonic 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 sell-off 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 selects loans initially, and a servicer monitors them later. We show that such

a separation of monitoring and screening tasks reduces both monitoring and screening efforts, 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.

Code Availability: The replication code is available in the Harvard Dataverse at doi.org/10.7910/DVN/HKZ2XA.

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

$$\max_{a_t \in [0,\bar{a}]} \left(a_t W_t - \frac{\phi a_t^2}{2} \right), \tag{A1}$$

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), that is, $a_t = \frac{W_t}{\phi}$, which is the first-order condition to (A1). The second-order condition to (A1), that is, $\frac{\partial^2}{\partial a_t^2}\left(a_t W_t - \frac{\phi a_t^2}{2} \right) = -\phi < 0$, is satisfied. Thus, the 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

$$\max_{q \in [0,\bar{q}]} \left(W_0(q) - \frac{\kappa q^2}{2} \right), \tag{A2}$$

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 (A2) satisfies $q > 0$.

Now observe that

$$q = \min\left\{ \frac{V_0(q)}{\kappa}, \bar{q} \right\} \tag{A3}$$

is the unique solution to (A2) if

$$\frac{\partial^2}{\partial q^2}\left(W_0(q) - \frac{\kappa q^2}{2} \right) = \frac{\partial}{\partial q} V_0(q) - \kappa < 0 \tag{A4}$$

holds for any $q \in [0, \bar{q}]$, in which case the objective in (A2) is strictly concave over the entire interval $[0, \bar{q}]$ and the first-order approach is valid. When optimal screening effort is interior, condition (A3) simplifies to (9), that is, $q = V_0/\kappa$, which is the first-order condition to (A2).

In what follows, we provide a sufficient condition for (A4) 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 (A4) can be rewritten as $Y_0(q) < \kappa$. Next, insert $a_t = W_t(q)/\phi$ into (12) to obtain

$$\dot{V}_t = \frac{dV_t(q)}{dt} = \left(\gamma + \Lambda - \frac{W_t(q)}{\phi} - q\right) V_t(q) - W_t(q), \tag{A5}$$

bearing in mind $\lambda_t = \Lambda - W_t(q)/\phi - q$. We now differentiate (A5) 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

$$Y_t(q) = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} \left(2V_s(q) + \frac{(V_s(q))^2}{\phi}\right) ds \tag{A6}$$

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

$$V_t(q) = \int_t^\infty e^{-\gamma(s-t) - \int_t^s \lambda_u du} W_s(q) ds \tag{A7}$$

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

$$\lambda_t = \Lambda - a_t - q \geq \Lambda - \bar{a} - \bar{q}. \tag{A8}$$

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

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

$$< \int_t^\infty e^{-(r+\Lambda-\bar{a}-\bar{q})(s-t)} 1 ds = \frac{1}{r + \Lambda - \bar{a} - \bar{q}} =: W^{max} \tag{A9}$$

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

Using these two relations (A8) and (A9) as well as (A7), we obtain that

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

$$\leq \frac{W^{max}}{\gamma + \Lambda - \bar{a} - \bar{q}} < \frac{1}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})} \tag{A10}$$

Using this inequality (A10) and the integral representation in (A6), we obtain that

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

$$\leq \int_t^\infty e^{-(\gamma+\Lambda-\bar{a}-\bar{q})(s-t)} \left(2V_s(q) + \frac{(V_s(q))^2}{\phi}\right) ds$$

$$< \frac{1}{(\gamma + \Lambda - \bar{a} - \bar{q})} \left(\frac{2}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})} + \frac{1}{\phi (r + \Lambda - \bar{a} - \bar{q})^2 (\gamma + \Lambda - \bar{a} - \bar{q})^2}\right).$$

As a result, a sufficient condition for (A4), that is, for

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

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

$$\kappa > \frac{2}{(r + \Lambda - \bar{a} - \bar{q})(\gamma + \Lambda - \bar{a} - \bar{q})^2} + \frac{1}{\phi (r + \Lambda - \bar{a} - \bar{q})^2 (\gamma + \Lambda - \bar{a} - \bar{q})^3}. \tag{A11}$$

That is, when (A11) holds, the first-order approach is valid and (A3) or, equivalently, (9) (due to $q < \bar{q}$) pins down screening effort. Note that (A11) is equivalent to condition (13) (Lemma 1). Also, notice that (13) is sufficient, but not necessary *per se*.

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

$$dP_t = (r+\lambda_t)P_t dt - 1dt + dC_t \tag{A12}$$

$$dW_t = (\gamma+\lambda_t)W_t dt + \frac{\phi a_t^2}{2}dt - dC_t. \tag{A13}$$

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

$$dF_t = dP_t + dW_t \tag{A14}$$

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

$$= (r+\lambda_t)(\underbrace{P_t + W_t}_{=F_t})dt - 1dt + \frac{\phi a_t^2}{2}dt - (\gamma - r)W_t dt. \tag{A15}$$

We can integrate (A14) over time, t, to get

$$F_t = \int_t^\infty e^{-r(s-t)-\int_t^s \lambda_u du}\left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s\right)ds, \tag{A16}$$

which is (20) from the main text.

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

$$F_0 - \frac{\kappa q^2}{2}, \tag{A17}$$

where F_0 is characterized in (A16). 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 the 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 subscripts (i.e., we write $F_t = F^B(q)$). In turn, the HJB equation simplifies to

$$rF^B(q) = \max_{W \in [0, F^B(q)], a \in [0,\bar{a}]} \left(1 - \frac{\phi a^2}{2} - (\gamma - r)W - \lambda F^B(q)\right) \tag{A18}$$

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

$$a^B(q) = \frac{F^B(q) - \phi(\gamma - r)}{\phi}, \tag{A19}$$

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 (A19) 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, that is, $\frac{\partial}{\partial q} a^B(q) \geq 0$.

Optimal monitoring effort implies the instantaneous default probability $\lambda = \lambda^B(q) = \Lambda - q - a^B(q)$. The law of motion (A12) and $dW_t = 0$ imply then that payouts to the agent take the form $dC_t = c^B(q)dt$ with

$$c^B(q) = (\gamma + \lambda^B(q))W^B(q) + \frac{\phi(a^B(q))^2}{2}. \tag{A20}$$

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

The objective (A17) can be rewritten as

$$\max_{q \in [0,\bar{q}]} \left(F^B(q) - \frac{\kappa q^2}{2}\right). \tag{A21}$$

At time $t = 0$, the agent chooses screening effort $q \in [0, \bar{q}]$ to maximize (A21), 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, that is, (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

$$\gamma W_t = \max_{a_t \in [0,\bar{a}]} \Big(-\underbrace{(\Lambda - a_t - q)}_{=\lambda_t} W_t - \frac{\phi a_t^2}{2} + c_t + \dot{W}_t \Big), \tag{A22}$$

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 (A22) under optimal a_t with respect to q so that

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

which is (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 $\frac{\partial c_t}{\partial q} = 0$.[30] We can integrate $\dot{V}_t = (\gamma + \lambda_t)V_t - W_t$ over time t 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 (A16):

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

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 the 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 a 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), that is, $\dot{V} = (\gamma + \lambda)V - W$. By the dynamic programming principle, total surplus $F(V)$ solves in any state V the HJB equation

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

---

30 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

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

which simplifies to (12), as desired.

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

$$a(V) = \frac{F(V) - F'(V)(V + \phi) - (\gamma - r)\phi}{\phi} \wedge \frac{F(V)}{\phi}. \tag{A23}$$

With slight abuse of notation, when the 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

$$V^B(q) = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q}. \tag{A24}$$

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

$$F(V^B(q)) = F^B(q). \tag{A25}$$

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 (A25)—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 $t = 0$

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

$$= F_0 + \ell(\kappa q - V_0). \tag{A26}$$

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 (A14) 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

$$F_t = \int_t^\infty e^{-r(s-t)} \left(1 - \frac{\phi a_s^2}{2} - (\gamma - r)W_s - \lambda_s F_s\right) ds. \tag{A27}$$

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

$$V_t = \int_t^\infty e^{-\gamma(s-t)} (W_s - \lambda_s V_s) ds. \tag{A28}$$

Using (A27) and (A28), we can rewrite the Lagrangian (A26) as

$$\mathcal{L}=\int_0^\infty e^{-rt}\left(1-(\gamma-r)W_t-\frac{\phi a_t^2}{2}-\lambda_t F_t\right)dt+\ell\left(\kappa q-\int_0^\infty e^{-\gamma t}(W_t-\lambda_t V_t)dt\right). \tag{A29}$$

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), that is, $a_t=W_t/\phi$. If interior, optimal effort a_t satisfies the first-order condition:

$$e^{-rt}(F_t-(\gamma-r)\phi-\phi a_t)-\ell e^{-\gamma t}(\phi+V_t)=0 \tag{A30}$$

Multiplying both sides of (A30) by e^{rt}, we obtain

$$F_t-(\gamma-r)\phi-\phi a_t-\ell e^{-(\gamma-r)t}(\phi+V_t)=0. \tag{A31}$$

Accounting for limited liability $W_t=\phi a_t\le F_t$ and $a_t\ge 0$, we can solve (A31) for

$$a_t=\max\left\{0,\frac{F_t-(\gamma-r)\phi-\ell e^{-(\gamma-r)t}(V_t+\phi)}{\phi}\right\}\ \wedge\ \frac{F_t}{\phi}. \tag{A32}$$

Taking the limit $t\to\infty$ in (A32) upon noticing that $\lim_{t\to\infty}W_t<\lim_{t\to\infty}F_t$ leads to

$$\lim_{t\to\infty}a_t=\lim_{t\to\infty}\left(\frac{F_t-(\gamma-r)\phi}{\phi}\right)<\lim_{t\to\infty}\frac{F_t}{\phi}, \tag{A33}$$

as V_t is bounded (see inequality (A10) in the proof of Lemma 1 and note that by definition, $V_t\ge 0$).

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 $\hat{a}$.[31] Note that by (A33),

$$\hat{a}=\max\left\{0,\frac{\hat{F}-(\gamma-r)\phi}{\phi}\right\}. \tag{A34}$$

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

$$\hat{F}=\frac{1-(\gamma-r)\phi\hat{a}-\frac{\phi\hat{a}^2}{2}}{r+\Lambda-\hat{a}-q}, \tag{A35}$$

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

$$\hat{a}=\arg\max_{a\in[0,\bar{a}]}\left(\frac{1-(\gamma-r)\phi a-\frac{\phi a^2}{2}}{r+\Lambda-a-q}\right), \tag{A36}$$

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 (A18). Now note that (A34) and (A35) as well as (A36) 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

$$\lim_{t\to\infty}V_t=\lim_{t\to\infty}\int_t^\infty e^{-\gamma(s-t)-\int_t^s\lambda_u du}W_s ds=\frac{\phi\hat{a}}{\gamma+\Lambda-\hat{a}-q}=V^B(q)\quad\text{and}\quad\lim_{t\to\infty}\dot{V}_t=0. \tag{A37}$$

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 (A37) would not hold.

[31] Equivalently,

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

That is, when $V_0=\kappa q > V^B(q)$, it follows that $\dot{V}_t<0$, with convergence according to (A37). Likewise, when $V_0=\kappa q < V^B(q)$, it follows that $\dot{V}_t>0$, with convergence according to (A37). 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 (A25), 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 (A23), $\lim_{V\to V^B(q)} a(V)=a^B(q)$ becomes equivalent to

$$\lim_{V\to V^B(q)} F'(V)=0, \tag{A38}$$

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

$$\lim_{V\to V^B(q)} F'(V)=\frac{F^B(q)-(\gamma-r)\phi}{\phi}\leq 0, \tag{A39}$$

so that $a(V)$ from (A23) 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

$$\max_{q\in[0,\bar{q}]}\left(F(V_0)-\frac{\kappa q^2}{2}\right) \quad \text{s.t.} \quad V_0=\kappa q. \tag{A40}$$

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 (A40) is

$$\frac{\partial F(V_0)}{\partial q}|_{q=q^*}+F'(V_0)\kappa=\kappa q^*, \tag{A41}$$

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

$$\kappa q^*=V_0\leq V^B(q^*)=\frac{W^B(q^*)}{\gamma+\Lambda-a^B(q^*)-q^*}, \tag{A42}$$

where the last equality follows (A24). 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

$$\kappa q^*=V_0\leq\frac{W^B(q^*)}{\gamma+\Lambda-a^B(q^*)-q^*}<\frac{F^B(q^*)}{r+\Lambda-a^B(q^*)-q^*}, \tag{A43}$$

where the first inequality follows (A42) 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, (A18) 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 (A18) with respect to q to obtain under the optimal controls $(W^B(q), a^B(q))$:

$$(r+\lambda)\frac{\partial F^B(q)}{\partial q} = F^B(q) \iff \frac{\partial F^B(q)}{\partial q} = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} > 0. \tag{A44}$$

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 $\frac{\partial^3 F^B(q)}{\partial q^3} > 0$.[32] Using (A44), we obtain

$$G'(q) = \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q. \tag{A45}$$

We also calculate

$$G''(q) = \frac{\partial^2}{\partial q^2} F^B(q) - \kappa \quad \text{and} \quad G'''(q) = \frac{\partial^3}{\partial q^3} F^B(q) > 0.$$

Because of $G'''(q) > 0$, 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 $q \in [0, q^B)$.

---

[32] To see this, note that $\frac{\partial a^B(q)}{\partial q} = \frac{\partial F^B(q)}{\partial q} \frac{1}{\phi}$. Thus, differentiating (A44) 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:

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

Next, define the (continuous) function of q:

$$K(q) := V^B(q) - \kappa q, \tag{A46}$$

with $V^B(q)$ from (A24), 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

$$K(q) = V^B(q) - \kappa q = \frac{W^B(q)}{\gamma + \Lambda - a^B(q) - q} - \kappa q < \frac{F^B(q)}{r + \Lambda - a^B(q) - q} - \kappa q = G'(q), \tag{A47}$$

where the first inequality uses that $r < \gamma$ and $W^B(q) \leq F^B(q)$ and the last equality uses (A45). 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 (A47).

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 by $\hat{q}$ 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 by $\hat{q}$, 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

$$\frac{\partial}{\partial q^*} \left(F^B(q^*) - \frac{\kappa (q^*)^2}{2} \right) = G'(q^*) > 0, \tag{A48}$$

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

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

$$F(V_0) - \frac{\kappa (q^\varepsilon)^2}{2} = F^B(q^\varepsilon) + F'(V^B(q^\varepsilon))\varepsilon + o(\varepsilon^2) - \frac{\kappa (q^\varepsilon)^2}{2} = F^B(q^\varepsilon) + o(\varepsilon^2) - \frac{\kappa (q^\varepsilon)^2}{2}$$

$$= \left(F^B(q^*) - \frac{\kappa (q^*)^2}{2} \right) + \frac{\partial}{\partial q^*} \left(F^B(q^*) - \frac{\kappa (q^*)^2}{2} \right) \varepsilon + o(\varepsilon^2),$$

which — by (A48) — 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 (A38)) 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 (A23) 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

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

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

Second, suppose that $F(V)>W(V)$ and the principal's limited liability constraint does not bind, and consider $V>V^B(q)$ 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

$$F''(V)=\frac{-(\gamma-r)F'(V)}{(\gamma+\lambda)V-W}. \tag{A49}$$

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 (A49) 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 (A39) implies $\lim_{V\to V^B(q)}F'(V)\leq 0$. Otherwise, when $a^B(q)>0$, then (A38) implies $F'(V^B(q))=0$ and — according to the expression for effort (A23):

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

owing to $\gamma>r$.

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

$$c_t = (\gamma + \lambda_t) W_t + \frac{\phi a_t^2}{2} - \dot{W}_t. \tag{A50}$$

If $F_t = W_t$, note that according to (A14), $\dot{F}_t = (\gamma + \lambda_t) F_t - 1 + \frac{\phi a_t^2}{2}$. Inserting the law of motion $\dot{F}_t = \dot{W}_t$ into (A50) 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 (A23):

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

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 (A50) implies $c_t > 0$.

A.4 Proof of Proposition 3 and Details about 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

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

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

$$\beta(V) - \beta'(V)\dot{V} L(V) = (\gamma + \Lambda - a(V) - q) W(V) + \frac{\phi a(V)^2}{2} - W'(V)\dot{V}, \tag{A51}$$

which is solved subject to

$$\lim_{V \to V^B(q)} \beta'(V) = 0 \iff \lim_{V \to V^B(q)} \beta(V) = c^B(q) = (\gamma + \Lambda - a^B(q) - q) W^B(q) + \frac{\phi (a^B(q))^2}{2}. \tag{A52}$$

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 (A51), as desired. Throughout, we assume the existence and uniqueness of a (nonconstant) solution to (A51) subject to (A52) 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 verify this conjecture. We calculate

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

$$=(1-\beta_t)\dot{L}_t-\dot{\beta}_t L_t=\frac{d}{dt}\left[(1-\beta_t)L_t\right],$$

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 $\dot{P}_t=\frac{d}{dt}\left[(1-\beta_t)L_t\right]$.

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 $(1-\beta_t)L_t\to P^B(q)$, that is, $\lim_{t\to\infty}P_t=\lim_{t\to\infty}(1-\beta_t)L_t$. Because, in addition, $P_t=(1-\beta_t)L_t$ implies $\dot{P}_t=\frac{d}{dt}\left[(1-\beta_t)L_t\right]$ 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 (A23) that the optimal effort $a(V)$ solves

$$a(V)=\frac{F(V)-F'(V)(V+\phi)-(\gamma-r)\phi}{\phi} \quad \wedge \quad \frac{F(V)}{\phi}, \tag{A53}$$

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, that is, $\hat{F}''(V)<0$.

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

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

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

Next, using (A53), we have

$$W(V)=\phi a(V)=\min\{F(V)-F'(V)(V+\phi)-(\gamma-r)\phi, F(V)\}. \tag{A54}$$

We can take the limit $\phi\to\infty$ for optimal continuation payoff in (A54), 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

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

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

Finally, we verify the 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. The second part of the proof then shows that $\lim_{t\to\infty}\beta_t = \beta^B = c^B = 0$ implies that the stake β_t reaches zero in *finite* time.

First, we recall that given q:

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

Next, the first-order derivative with respect to a satisfies

$$\frac{\partial(\Lambda+r)F^B}{\partial a} = F^B-(\gamma-r)\phi-\phi a \iff \frac{\partial F^B}{\partial a} = \frac{F^B-(\gamma-r)\phi-\phi a}{\Lambda+r}. \tag{A55}$$

Provided $a<\bar{a}$, we have $a=a^B=0$ only if

$$\phi \geq \frac{1}{(r+\Lambda-q)(\gamma-r)} \tag{A56}$$

Note that this necessary condition depends on the level of screening effort, which is endogenous. 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 (A55))

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

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

We now proceed by deriving a sufficient condition for $a=a^B=0$, which only depends on exogenous model parameters. Because $\frac{\partial F^B}{\partial q}>0$, we obtain

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

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

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

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

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

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

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

where the second equality uses (A55), 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

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

that is, 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 $\lim_{t\to\infty}c_t=0$.

A.6.2 Part 2. It remains to check whether V_t reaches 0 in a 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 $\lim_{V\to0}\dot{V}=0$. As $\lim_{V\to0}F(V)=F^B(q)$, we have

$$F'(0):=\lim_{V\to0}F'(V)=\frac{F^B(q)-(\gamma-r)\phi}{\phi}<0,$$

so that $a(V)$ from (A23) 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

$$F''(V)=\frac{-(\gamma-r)F'(V)}{\dot{V}}. \tag{A57}$$

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 (A57). 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

$$0 \le \lim_{V\to 0} \frac{-V^{-\alpha}}{G(V)} = \lim_{V\to 0} \frac{-\alpha V^{\alpha-1}}{G'(V)} = \lim_{V\to 0} \frac{\alpha V^{\alpha-1}}{\frac{(\gamma-r)F'(V)}{G(V)}} = \lim_{V\to 0} \frac{-\alpha V^{\alpha-1} G(V)}{(\gamma - r)F'(V)} = \lim_{V\to 0} \frac{-\alpha V^{\alpha-1} G(V)}{(\gamma - r)F'(0)}$$

$$= \lim_{V\to 0} \frac{\alpha V^{\alpha-1} G(V)}{(\gamma - r)F'(0)} \le \lim_{V\to 0} \frac{\alpha V^{\alpha-1}(-V)}{(\gamma - r)F'(0)} = \lim_{V\to 0} \frac{\alpha V^{\alpha}}{-(\gamma - r)F'(0)} = 0,$$

where we used L'Hopital's rule in the first equality and that $G(V) < -V$ in a neighborhood 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 $T = \inf\{t \ge 0 : V_t = \epsilon\}$. As $\epsilon > 0$ and $G(V) < 0$ for $V \ge \epsilon$, it readily follows that T is finite, that is, $T < \infty$.

Next, for times $t \ge T$, consider the ODE $X_t = -X_t^\alpha$ with $X_T = K > 0$ for $\alpha \in (0,1)$ which admits the general solution:[33]

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

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 $X_T = V_T = \epsilon$. Because of $G(V) \le -V^\alpha$ as well as $\dot{V}_t = G(V_t)$ and $\dot{X}_t = -X_t^\alpha$, it follows that $V_t \le X_t$ for $t \ge T$. As such, V_t reaches 0 in finite time $T'' \le T' < \infty$. Thus, in the implementation, β_t reaches $\beta(0) = \beta^B = 0$ in finite time, which was to be shown.

---

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

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

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* 132:1–25.

Begley, T. A., and A. Purnanandam. 2016. Design of financial securities: Empirical evidence from private-label rmbs deals. *Review of Financial Studies* 30:120–61.

Benmelech, E., J. Dlugosz, and V. Ivashina. 2012. Securitization without adverse selection: The case of CLOs. *Journal of Financial Economics* 139:452–77.

Bernstein, S., X. Giroud, and R. R. Townsend. 2016. The impact of venture capital monitoring. *Journal of Finance* 71:1591–622.

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* 74:345–90.

Blickle, K., Q. Fleckenstein, S. Hillenbrand, and A. Saunders. 2022. The myth of the lead arranger's share. Working paper, New York University.

Blickle, K., C. Parlatore, and A. Saunders. 2023. Specialization in banking. Staff reports, Federal Reserve Bank of New York.

Bord, V. M., and J. A. Santos. 2015. Does securitization of corporate loans lead to riskier lending? *Journal of Money, Credit and Banking* 47:415–44.

Bruche, M., F. Malherbe, and R. R. Meisenzahl. 2020. Pipeline risk in leveraged loan syndication. *Review of Financial Studies* 33:5660–705.

Chen, H., Y. Xu, and J. Yang. 2021. Systematic risk, debt maturity, and the term structure of credit spreads. *Journal of Financial Economics* 139:770–99.

Chen, M., S. J. Lee, D. Neuhann, and F. Saidi. 2023. Less bank regulation, more non-bank lending. Working Paper, Federal Reserve Bank of Boston.

Daley, B., and B. Green. 2012. Waiting for news in the market for lemons. *Econometrica* 80:1433–504.

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

DeMarzo, P., and Z. He. 2021. Leverage dynamics without commitment. *The Journal of Finance* 76:1195–250.

DeMarzo, P. M., and Y. Sannikov. 2006. Optimal security design and dynamic capital structure in a continuous-time agency model. *Journal of Finance* 61:2681–724.

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* 25:3217–58.

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

Drucker, S., and M. Puri. 2009. On loan sales, loan contracting, and lending relationships. *Review of Financial Studies* 22:2835–72.

Gopalan, R., V. Nanda, and V. Yerramilli. 2011. Does poor performance damage the reputation of financial intermediaries? evidence from the loan syndication market. *Journal of Finance* 66:2083–120.

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

Gottardi, P., H. Moreira, and W. Fuchs. 2022. Time trumps quantity in the market for lemons. Working Paper, University of Texas at Austin.

Gryglewicz, S., and S. Mayer. 2022. Dynamic contracting with intermediation: Operational, governance, and financial engineering. *Journal of Finance* 78:2779–836.

Gustafson, M., I. Ivanov, and R. Meisenzahl. 2021. Bank monitoring: Evidence from syndicated loans. *Journal of Financial Economics* 139:91–113.

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

Haque, S., S. Mayer, and T. Wang. 2023. How private equity fuels non-bank lending. Working Paper, Board of Governors of the Federal Reserve System.

Hartman-Glaser, B. 2017. Reputation and signaling in asset sales. *Journal of Financial Economics* 125:245–65.

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

Heitz, A. R., C. Martin, and A. Ufier. 2022. Bank monitoring with on-site inspections. Research Paper, FDIC Center for Financial Research.

Hoffmann, F., R. Inderst, and M. M. Opp. 2021. Only time will tell: A theory of deferred compensation. *Review of Economic Studies* 88:1253–78.

———. 2022. The economics of deferral and clawback requirements. *Journal of Finance* 77:2423–70.

Holmstrom, B. 1989. Agency costs and innovation. *Journal of Economic Behavior & Organization* 12:305–27.

Hu, Y., and F. Varas. 2021. Intermediary financing without commitment. Working Paper, The University of North Carolina.

Irani, R. M., R. Iyer, R. R. Meisenzahl, and J.-L. Peydro. 2021. The rise of shadow banking: Evidence from capital regulation. *Review of Financial Studies* 34:2181–235.

Irani, R. M., and R. R. Meisenzahl. 2017. Loan sales and bank liquidity management: Evidence from a us credit register. *Review of Financial Studies* 30:3455–501.

Ivashina, V. 2009. Asymmetric information effects on loan spreads. *Journal of Financial Economics* 92:300–19.

Jiang, S., S. Kundu, and D. Xu. 2023. Monitoring with small stakes. Working Paper, University of Florida.

Keys, B. J., T. Mukherjee, A. Seru, and V. Vig. 2010. Did securitization lead to lax screening? evidence from subprime loans. *Quarterly Journal of Economics* 125:307–62.

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* 138:105550–.

Malamud, S., H. Rui, and A. Whinston. 2013. Optimal incentives and securitization of defaultable assets. *Journal of Financial Economics* 107:111–35.

Malenko, A. 2019. Optimal dynamic capital budgeting. *Review of Economic Studies* 86:1747–78.

Nadauld, T. D., and M. S. Weisbach. 2012. Did securitization affect the cost of corporate debt? *Journal of Financial Economics* 105:332–52.

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* 63:1291–314.

Piskorski, T., and M. M. Westerfield. 2016. Optimal dynamic contracts with moral hazard and costly monitoring. *Journal of Economic Theory* 166:242–81.

Purnanandam, A. 2011. Originate-to-distribute model and the subprime mortgage crisis. *Review of Financial Studies* 24:1881–915.

Saunders, A., A. Spina, S. Steffen, and D. Streitz. 2021. Corporate loan spreads and economic activity. Working Paper, New York University.

Sufi, A. 2007. Information asymmetry and financing arrangements: Evidence from syndicated loans. *Journal of Finance* 62:629–68.

Varas, F., I. Marinovic, and A. Skrzypacz. 2020. Random inspections and periodic reviews: Optimal dynamic monitoring. *Review of Economic Studies* 87:2893–937.

Wang, Y., and H. Xia. 2014. Do lenders still monitor when they can securitize loans? *Review of Financial Studies* 27:2354–91.