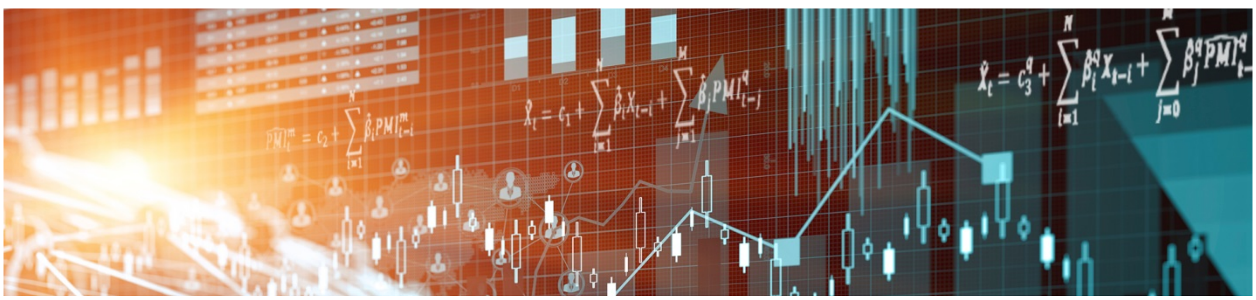
Staff Working Paper/Document de travail du personnel 2016-56

Bank Screening Heterogeneity



by Thibaut Duprey

www.bank-banque-canada.ca

Bank of Canada Staff Working Paper 2016-56

December 2016

Bank Screening Heterogeneity

by

Thibaut Duprey

Financial Stability Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
tduprey@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

This paper benefited greatly from discussions with Victor Aguirregabiria, Jason Allen, Brunos Biais, Régis Breton, Edouard Challe, Gabriel Chodorow Reich, Laurent Clerc, Nicolas Coeurdacier, Evren Damar, Jason Donaldson, Douglas Gale, Itay Goldstein, Jean Imbs, Luigi Iovino, Enisse Kharroubi, Robert Marquez, David Martinez-Miera, Héctor Pérez Saiz, Guillaume Plantin, Jérôme Pouyet, Xavier Ragot and Liyan Yang. Earlier versions of this work also benefited from comments by seminar participants at the Paris School of Economics, the Banque de France, the Bank of Canada, the 2013 French Economic Association Congress, the 2013 European Economic Association Congress, the 2014 Royal Economic Society Congress and the 2014 Winter Conference of the Multinational Finance Society. I gratefully acknowledge financial support from the Banque de France and the Association Nationale de la Recherche et de la Technologie. The opinions expressed herein do not reflect the position of the Banque de France or the Bank of Canada. Any remaining errors are mine.

Abstract

Production efficiency and financial stability do not necessarily go hand in hand. With heterogeneity in banks' abilities to screen borrowers, the market for loans becomes segmented and a self-competition mechanism arises. When heterogeneity increases, the intensive and extensive margins have opposite effects. Bank informational rents unambiguously decrease welfare and distort effort incentives. But the bank most efficient at screening expands its market share by competing against itself to offer effort-inducing contracts, which decreases the share of non-performing loans. A macroprudential authority acting alone reinforces this tension. Optimality is restored by targeting lending policies toward borrowers with intermediate abilities.

Bank topics: Financial institutions; Financial stability; Financial system regulation and policies
JEL codes: G14; G21; L13

Résumé

L'efficience de la production et la stabilité financière ne vont pas nécessairement de pair. Dans un contexte d'hétérogénéité entre les banques quant à leur capacité à sélectionner les entrepreneurs, il s'opère une segmentation du marché du crédit et un phénomène d'autoconcurrence apparaît. Lorsque l'hétérogénéité s'accroît, les marges intensive et extensive ont des effets opposés. Les rentes informationnelles dont bénéficient les banques diminuent sans ambiguïté le bien-être et modifient les incitations à l'effort. La banque la plus efficace en matière de sélection des entrepreneurs voit cependant sa part de marché progresser lorsqu'elle se fait concurrence à elle-même en offrant des contrats qui incitent les entrepreneurs à intensifier leurs efforts, ce qui abaisse la proportion de prêts non productifs. Une autorité macroprudentielle agissant seule renforce cette dynamique. Le rétablissement de l'optimum nécessite des mesures favorisant le crédit aux entrepreneurs ayant une capacité de réussite moyenne.

Sujets : Institutions financières, Stabilité financière, Réglementation et politiques relatives au système financier
Codes JEL : G14; G21; L13

Non-Technical Summary

This paper investigates the financial stability impact of banks' heterogeneous access to borrowers' information. When banks have different abilities to screen borrowers, the bank with the best technology has a comparative advantage, and one could think that this bank would serve the whole market. However, the paper shows that other, less efficient banks can survive in equilibrium by specializing in financing projects where borrowers find themselves better off if they do not reveal information.

This set-up creates a tension between the loan size offered to borrowers and the market share of the banks. Production efficiency may come at the cost of financial stability. When bank screening heterogeneity increases, the most efficient bank increases its informational rent, and this unambiguously decreases production efficiency. Screened borrowers face less attractive loan contracts; fewer borrowers are screened, and they are credit-rationed, so, in the aggregate, the share of performing loans decreases.

However, the additional rent extraction of the most efficient bank allows it to expand its market share. This bank can either compete more to gain additional customers, or strategically reduce competition on other loan segments to make screened lending relatively more attractive. This tends to increase the share of performing loans. This tension can also be thought of in the context of relationship banks versus non-banks, entry into less efficient banking markets, or regulation-implied comparative advantages.

Several implications arise. First, the impact of bank competition on banking risks is non-linear due to market segmentation. Second, the borrowers that suffer most from increased banking heterogeneity are those with intermediate abilities. Third, a macroprudential authority averse to excessive loan losses can successfully restore financial stability by applying a bank-specific tax scheme, but this harms production efficiency as it further increases banking heterogeneity. Last, allocation efficiency can be restored if a fiscal authority subsidizes the borrowers that suffer most from banks' rent extraction.

1 Introduction

One of banks' key roles is to channel funds efficiently via information acquisition about borrowers' type. One could think that a single institution with the best information production technology would dominate the market for loans. Yet, some degree of information-driven market segmentation can allow for the coexistence of several banks with different risk profiles.[1] This paper focuses on the financial stability impact of market segmentation generated by banks' heterogeneous access to information.

This paper focuses on heterogeneity in screening technologies: entrepreneurs with unobserved effort costs (first inefficiency) can be screened, but some banks have a comparative advantage in screening costs (second inefficiency). Thus, different loan contracts attract entrepreneurs with different costs of effort, so screening occurs in a context of adverse selection. The credit market splits into two segments: one in which loan contracts embed screening and only entrepreneurs with costs of effort low enough self-select that type of contract, and one in which banks extend credit indiscriminately.[2] I find that variations in bank screening heterogeneity in the cross-section can generate a financial stability/production efficiency trade-off. Banking heterogeneity unambiguously decreases welfare away from the optimum. However, banking heterogeneity can distort incentives towards a lower economy-wide share of non-performing loans closer to the optimum.

The model is composed of three stages. First banks compete to offer lending rates that maximize the profits. The screening decision for each loan is binary, so banks can decide to offer two rates, with and without screening. Then loan volumes are determined by loan demand, while still ensuring incentives are met given the offered interest rate. Banks can always prevent opportunistic behaviours of entrepreneurs by rationing credit. Last, entrepreneurs decide which project to undertake to maximize profits, given the

[1] Many alternative reasons for information-related market segmentation can be thought of: information could be multi-dimensional (Petersen, 2004), transaction and relationship lenders could require different levels of informations (Boot and Thakor, 2000), large banks could be less efficient at creating lending relationships (Stein, 2002), markets could be geographically segmented (Dick and Lehnert, 2010), or regulation could create informational advantages (Hakenes and Schnabel, 2011).

[2] Sorting of borrowers can arise without heterogeneity in bank screening cost. In Bester (1985), borrowers with a higher probability of default accept contracts with higher interest rates in order to save on the cost of collateral. In the context of capital accumulation, Cetorelli and Peretto (2000) show that banks screen entrepreneurs only with some probability, thereby screened and unscreened loans coexist.

set of loan contracts they face. If entrepreneurs pick a loan contract with screening, their unobserved cost of effort is revealed to the bank; the contractual agreement is now conditional on costly effort by entrepreneurs who are incentivized to undertake a project with a higher probability of success.

The mechanism for the result is as follows. On the one hand, heterogeneity generates an informational rent for the bank most efficient at screening.[3] This is equivalent to rationing credit to effort-making entrepreneurs with higher probabilities of success. When banking heterogeneity increases, limit-pricing competition on the intensive margin increases the economy-wide share of non-performing loans. On the other hand, endogenous market segmentation arises, with the most efficient bank making profits by being the only one providing screening. This bank seeks to increase its market share by (i) directly competing with itself to offer more attractive loan contracts to the marginal entrepreneur, and (ii) indirectly competing with itself to make alternative projects less attractive. It can do so by strategically reducing competition with other banks on alternative market segments.[4] When banking heterogeneity increases, self-competition on the extensive margin allows expanding the market share of screened entrepreneurs, and this decreases the economy-wide share of non-performing loans.

The results have several implications. First, the type of bank competition is important. If banks cannot extract rents, the comparative advantage in screening has no impact on loan pricing. Conversely, when the banking sector is concentrated (duopoly), all banks extract rents. Entry improves allocation efficiency, but the impact of bank competition on banking risks is non-linear (as opposed to Boyd and De Nicolo, 2005) and depends on the relative competition across the different market segments.

The model delivers a specific market segmentation. Entrepreneurs with low costs of providing effort are more likely to be subject to screening: they have more to gain by

[3] Hauswald and Marquez (2003) show that better abilities to process information by some banks widen the informational gap compared with banks that do not invest in information acquisition. Informed banks extract a larger rent and charge a higher rate, but this is capped by the possibility of informational spillovers. Empirically, Schenone (2010) shows that relationship banks that gain specific information can charge a higher rate as long as this information remains proprietary.

[4] This mechanism relates to the two reasons for strategic information acquisition by banks in a model where information depends on the distance to the borrower: reduce bank competition and extend market shares (Hauswald and Marquez, 2006).

self-selecting a bank that learns how good they are. Conversely, with ex post monitoring to prevent opportunistic behaviours during the realization of a project, the high-quality entrepreneurs enter a contract with a bank that provides less monitoring, as they do not require much monitoring to behave.[5]

The model delivers another testable hypothesis: entrepreneurs with intermediate effort costs bear the largest premium over the fair pricing of loans. On the one hand, entrepreneurs with lower costs of effort are more productive and thus face stronger competition by banks. On the other hand, entrepreneurs with higher costs of effort are less productive and more likely to change their production plans. To retain its market share and preserve incentives, the bank has to offer them more attractive loan contracts.

In the presence of banking heterogeneity, a macroprudential authority averse to excessive loan losses above the optimum can successfully restore financial stability by applying a bank-specific tax scheme targeted towards riskier lenders. However, a heterogeneous levy on banks reinforces the comparative advantage of the bank most efficient in screening, which harms allocation efficiency. Allocation efficiency can be restored if a fiscal authority subsidizes the borrowers that suffer most from banks' rent extraction.

Last, an interesting by-product of banking heterogeneity is that it creates heterogeneous lending cycles. If moral hazard tightens in bad times, loans for riskier projects are more cyclical (Berger and Udell, 2004; Dell'Ariccia et al., 2008), and aggregate lending by banks that do not screen is also more cyclical (Beck et al., 2015).

The rest of the paper is structured as follows. The next section presents the set-up of the model. Section 3 derives the equilibrium. Section 4 presents the distortions associated with banking heterogeneity and describes the impact on aggregate banking risks. Section 5 extends the discussion to regulation and alternative situations. Section 6 concludes.

[5]See, for example, theoretical papers such as Boot and Thakor (2000) or Martinez-Miera and Repullo (2015) and empirical papers such as Petersen and Rajan (1995) in the context of relationship lending.

2 Model set-up

2.1 Heterogeneous entrepreneurship sector

I consider a continuum e of entrepreneur type, where a low e is associated with a low cost of providing effort. For simplicity, the cumulative density function $F(e)$ is assumed to follow a uniform distribution $e \to \text{U}\,[e^{min}; e^{max}]$. Entrepreneurs are cashless and, to undertake project j, they must borrow all the funds $B_{j,n}$ from a single financial intermediary, n. Entrepreneurs have a constant return-to-scale production function, $\alpha B_{j,n}$.

Given their type e, entrepreneurs balance their effort provision against risk taking. Entrepreneurs can choose between three risk profiles $j \in \{H, M, L\}$ associated with a high, medium or low probability of success, respectively p_H, p_M and p_L, with $1 > p_H > p_M > p_L > 0$. If entrepreneurs decide to gamble, their project matures with probability p_L, which is not sufficient for the project to be socially efficient, but they obtain a private benefit b identical across all agents and increasing with project size. This is the standard source of moral hazard in Holmstrom and Tirole (1997). Entrepreneurs face a binary effort decision, which adds another layer of moral hazard. If entrepreneurs do not provide effort, their project is successful with probability p_M. If entrepreneurs provide effort, their project is successful with a higher probability p_H, but they have to bear the cost of effort $eB_{j,n}^2/2$ increasing with project size.[6]

For a given productivity level α, risk-free interest rate r (normalized to 1), entrepreneur type e and bank screening costs s, both H and M projects have a positive marginal NPV, while project L has a negative marginal net present value (NPV):

Assumption 1. $\alpha p_H - r - eB_{j,n} - s > \alpha p_M - r > 0 > \alpha p_L - r$ *and* $r = 1$.

I further assume that everything is lost in the case of default, and default is uncorrelated across entrepreneurs. Henceforth, the funds $B_{j,n}$ needed by an entrepreneur to run the firm have to be borrowed at a rate $R_{j,n}$ higher than the market risk-free interest rate r.

[6] The cost of effort does not need to increase with the square of the project size. This assumption simply adds convexity to simplify the exposition of the results.

The expected profit π_j of an entrepreneur is expressed as the expected return of the project j it undertakes, net of the expected repayment of the loan contract signed with bank n and net of the cost/gain associated with each risk profile:

$$\underset{j \in \{H,M,L\}}{\text{Max}} \pi_j = p_j(\alpha - R_{j,n})B_{j,n} + \Psi_j \frac{B_{j,n}^2}{2} \quad \text{where} \quad \Psi_j = \begin{cases} -e, & \text{for j=H} \\ 0, & \text{for j=M} \\ b, & \text{for j=L} \end{cases} . \tag{1}$$

2.2 The heterogeneous banking sector

Banks know the private benefit b of entrepreneurs from gambling. Following for example Diamond (1984) and Fama (1985), the rationale for having banks is their access to a superior technology compared with outside financiers that allows them to perfectly screen out negative NPV projects $j = L$. However, banks ignore the individual type e of the entrepreneurs and only know the distribution of effort types among the pool of entrepreneurs. The choice of the banks is to decide to offer loan contracts with or without screening. The type e of the entrepreneur is revealed if the bank screens.

Heterogeneity in the banking sector is introduced in the form of differences in the cost of screening. For simplicity, I assume that banks do not invest in technology to close this gap.[7] I consider N banks with different costs of screening. Bank $n = 1$ has the lowest screening costs, $s_1 < ... < s_n < ... < s_N$.

Each competing bank designs several loan contracts $C_{j,n}(R_{j,n}, B_{j,n})$ given by the interest rate $R_{j,n}$ and the loan size $B_{j,n}$ that cover the cost of funding r, and possibly the screening cost s_n if it decides to acquire information. The expected profit Π_n for banks $n \in \{1, ..., N\}$ is as follows:

$$\underset{\{R_{j,n}, B_{j,n}\}}{\text{Max}} \Pi_n = p_j R_{j,n} B_{j,n} - s_n - r B_{j,n} \ . \tag{2}$$

[7] The literature on the "X-efficiency" suggests that production may not always lie on the outer-bound production possibility frontier and that inefficiencies can persist, for instance if competition is not perfect or if large disutilities of effort exist. Hughes and Mester (2009) review the large literature that looks at how banks' technologies deviate from the optimal production frontier to explain banks' comparative advantages in producing informational-intensive assets.

I do not consider differences on the liability side in order to isolate the effect of bank screening cost heterogeneity. Despite the fact that one type of bank is clearly dominant—higher screening costs are not compensated for instance by easier access to capital markets—it has non trivial implications in a model with heterogeneity among entrepreneurs.

2.3 Equilibrium

A bank loan contract $C_{j,n}$ offered by bank n to finance project j is defined by a loan supply schedule $B_{j,n}$ pinned down by moral hazard and an interest rate $R_{j,n}$ pinned down by bank competition. The subgame-perfect Nash equilibrium is defined by:

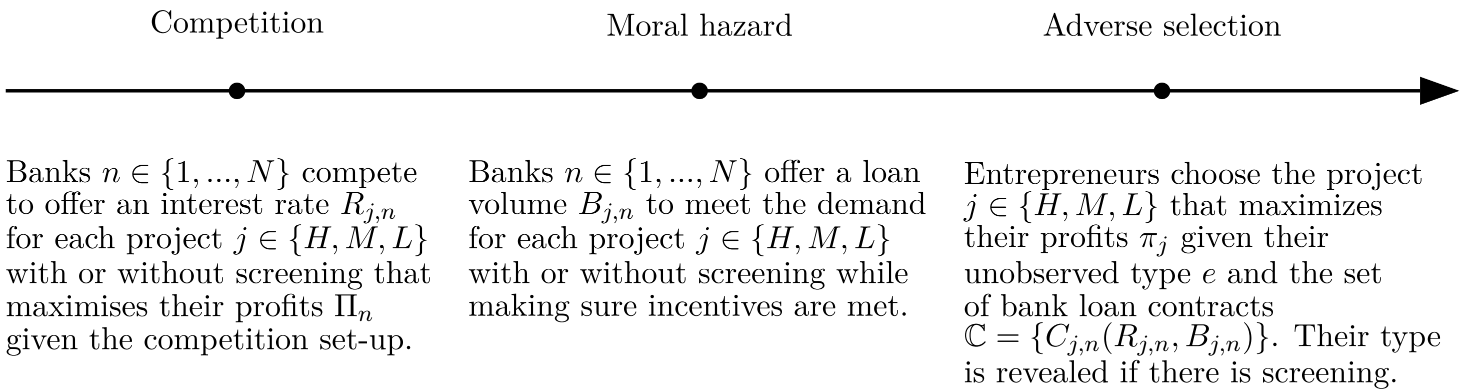
- a set of contracts $\mathbb{C}$ with and without screening $\mathbb{C} = \{C_{j,n}(R_{j,n}, B_{j,n})\}_{j\in\{H,M\},n\in\{1,...,N\}}$ given bank screening technology s_n and entrepreneur type e;
- entrepreneurs' preferences over the set of contracts summarized by the cut-off $\bar{e} \subset [e_{min}, e_{max}]$ for the entrepreneur indifferent between projects $j \in \{H, M\}$, obtained via the adverse selection of bank loan contracts.

Table 1 and Figure 1 display the matrix of payoffs and the sequence of events.

Table 1: Matrix of payoffs

for a bank $n \in \{1, ..., N\}$ / for an entrepreneur $e \in [e^{min}; e^{max}]$	screening contract $C_{H,n}\left(R_{H,n}, B_{H,n}\right)$	non-screening contract $C_{M,n}\left(R_{M,n}, B_{M,n}\right)$
High yield/effort	$p_H R_{H,n} B_{H,n} - s_n - rB_{H,n}$ $p_H \alpha B_{H,n} - p_H R_{H,n} B_{H,n} - eB_{H,n}^2/2$	$p_H R_{M,n} B_{M,n} - rB_{M,n}$ $p_H \alpha B_{M,n} - p_H R_{M,n} B_{M,n} - eB_{M,n}^2/2$
Medium yield/no effort	$p_M R_{H,n} B_{H,n} - s_n - rB_{H,n}$ $p_M \alpha B_{H,n} - p_M R_{H,n} B_{H,n}$	$p_M R_{M,n} B_{M,n} - rB_{M,n}$ $p_M \alpha B_{M,n} - p_M R_{M,n} B_{M,n}$
Low yield/gambling	$p_L R_{H,n} B_{H,n} - s_n - rB_{H,n}$ $p_L \alpha B_{H,n} - p_L R_{H,n} B_{H,n} + bB_{H,n}^2/2$	$p_L R_{M,n} B_{M,n} - rB_{M,n}$ $p_L \alpha B_{M,n} - p_L R_{M,n} B_{M,n} + bB_{M,n}^2/2$

Figure 1: Timing of the model



3 Model with bank screening heterogenity

The equilibrium is obtained by solving the model backward with three stages.

3.1 Choice of project by entrepreneurs

Entrepreneurs can choose between projects $j \in \{H, M, L\}$. Project L has negative NPV, and banks design incentives to avoid it. Given their own type e, and depending on the set of loan contracts $\mathbb{C}$ offered by the banks, entrepreneurs can decide to undertake project H or M. For each project type, one bank contract is (possibly weakly) superior. The choice of entrepreneurs' risk profile is thus an adverse selection process. The entrepreneur $\bar{e}_1$ indifferent between undertaking an H or an M project is given by the following equation holding as an equality:

$$\begin{aligned} C_H\left(R_H, B_H\right) &\succ C_M\left(R_M, B_M\right) \\ p_H\left(\alpha - R_H\right) B_H - eB_H^2/2 &> p_M\left(\alpha - R_M\right) B_M \\ e &< \bar{e}_1\left(R_H, B_H, R_M, B_M\right) \ . \end{aligned} \tag{3}$$

This threshold $\bar{e}_1$ generates endogenous market segmentation.

Entrepreneurs could also pick the contract without screening $C_M\left(R_M, B_M\right)$ that is not conditional on the entrepreneur's type, but still undertake the H project by providing effort without the bank knowing it. However, it seems reasonable to assume that an entrepreneur who is willing to provide effort is better off if the bank knows his type so that he obtains better lending conditions. This is equivalent to:

$$p_H\left(\alpha - R_H\right) B_H - eB_H^2/2 > p_M\left(\alpha - R_M\right) B_M > p_H\left(\alpha - R_M\right) B_M - eB_M^2/2 \ .$$

This assumption corresponds to a lower bound on the distribution of effort abilities.

Assumption 2. $e_{min} > \frac{2\Delta_H(\alpha - R_M)}{B_M}$.

3.2 Choice of loan supply

Banks know that entrepreneurs are heterogeneous and will behave differently depending on their type above or below $\bar{e}_1$. Loan supply is then conditional on banks screening to learn the type e of the entrepreneur. Note that, conditional on an interest rate that satisfies a bank's participation constraint, bank profits increase in the loan size offered to entrepreneurs.

If entrepreneurs undertake the H project. Upon undertaking project H if $e < \bar{e}_1$, an entrepreneur e asks for the loan size that maximizes his profits, and the bank provides the loan conditional on incentives being met:

$$\underset{B_H}{\text{Max}} \qquad p_H(\alpha - R_H)B_H - eB_H^2/2 \tag{4}$$

$$\text{subject to:}$$

$$\alpha > R_H \tag{PC}$$

$$e < \bar{e}_1 \tag{AS}$$

$$p_H(\alpha - R_H)B_H - eB_H^2/2 > p_M(\alpha - R_H)B_H \tag{EC}$$

$$p_H(\alpha - R_H)B_H - eB_H^2/2 > p_L(\alpha - R_H)B_H + bB_H^2/2 \tag{IC}$$

$$R_H \quad \text{given.}$$

The first constraint is the entrepreneur's participation constraint (PC). The second constraint comes from the adverse selection (AS) of loan contracts. The two subsequent constraints refer to the two sources of moral hazard in the model. The effort compatibility (EC) constraint ensures the entrepreneur provides costly effort. The incentive compatibility (IC) constraint ensures the entrepreneur does not gamble to reap a private benefit. If effort is too costly for a larger project size, or if constraints EC or IC are binding, the loan supply schedule is given by:

$$B_H = \min\left\{\frac{p_H(\alpha - R_H)}{e}; \frac{2\Delta_H(\alpha - R_H)}{e}; \frac{(p_H - p_L)(\alpha - R_H)}{e + b}\right\} . \tag{5}$$

I denote $\Delta_H = p_H - p_M$. The higher the effort cost or the interest rate, the smaller the size of the loan. There is an interior solution given by $B_H = \frac{p_H(\alpha - R_H)}{e}$ under the conditions that $2p_M < p_H$ and $b < e\left(1 - \frac{2p_L}{p_H}\right)$.

If entrepreneurs undertake the M project. Upon undertaking project M if $e > \bar{e}_1$, an entrepreneur e asks for the largest possible loan size that is still incentives-compatible:

$$\begin{aligned} \underset{B_M}{\text{Max}} \qquad & p_M(\alpha - R_M)B_M && (6)\\ \text{subject to:} & \\ \alpha > R_M & && \text{(PC)}\\ e > \bar{e}_1 & && \text{(AS)}\\ p_H\left(\alpha - R_M\right)B_M - eB_M^2/2 < p_M\left(\alpha - R_M\right)B_M & && \text{(EC)}\\ p_M\left(\alpha - R_M\right)B_M > p_L\left(\alpha - R_M\right)B_M + bB_M^2/2 & && \text{(IC)}\\ R_M \quad \text{given.} & \end{aligned}$$

The first constraint is the entrepreneur's PC. The second constraint comes from the AS of loan contracts and ensures the loan is offered to entrepreneurs that are better off undertaking project M. The EC constraint should not hold to ensure the entrepreneur does not switch to project H. The fourth IC constraint ensures the entrepreneur does not gamble. The loan supply is given by the IC constraint that is binding in equilibrium:

$$B_M = \frac{2\Delta_M\left(\alpha - R_M\right)}{b} \ . \tag{7}$$

I denote $\Delta_M = p_M - p_L$. The higher the private benefit of gambling or the interest rate, the smaller the size of the loan. Under assumption 2, the EC constraint is not binding, and assumption 2 is rewritten in terms of model parameters $e_{min} > b\frac{\Delta_H}{\Delta_M}$. It means that the minimal cost of moral hazard associated with deviating to project H, adjusted for the change in probability of success, is larger than the gain of moral hazard associated with deviating to project L, adjusted for the change in probability of success. So the possibility of gambling with project L is binding earlier than deviating to project H.[8]

3.3 Choice of interest rate and screening decision

Conditional on entrepreneurs' choice (adverse selection) and conditional on the loan supply (capped by moral hazard), banks decide on the offered interest rate. Banks make sure their own PC holds, together with the one of entrepreneurs ($\alpha > R_j$) that gives an upper bound on the offered rate. The interest rate is pinned down by the intensity of bank competition, that is to say the ability of banks to compete à la Bertrand given their cost of providing loans.

Competition on the intensive margin. If a bank n decides to screen, it learns the type e of the entrepreneur and is able to provide a loan $B_{H,n}$ that is conditional on e. However, the bank incurs a cost s_n, and its PC is given by:

$$p_H R_{H,n} B_{H,n} - s_n \quad \geq \quad r B_{H,n} \ . \tag{8}$$

Banks play Bertrand price competition, so with heterogeneous screening costs s_n, the equilibrium interest rate $R_{H,n}$ is given by limit pricing where the bank most efficient at screening can drive its competitors out of the market segment with screening. The bank most efficient at screening $n = 1$ with screening cost $s_1 < s_2$ sets the interest rate that ensures zero profits $\Pi_2 = 0$ of the bank with the second-best screening technology s_2. Equation (8) holds as an equality for s_2, given the loan supply of equation (5). Assuming an interior solution for the loan supply schedule, and recognizing that the smallest root is the solution given the competition in prices, the interest rate is given by:

$$R_{H,1} = \frac{1}{2p_H}\left(r + p_H\alpha - \sqrt{(r + p_H\alpha)^2 - 4(rp_H\alpha + s_2 e)}\right) \ . \tag{9}$$

[8]This allows keeping the model simple. The only thing required for the results is that some entrepreneurs will decide to accept the loan contract for M projects without providing effort, that is $e_{max} > b\frac{\Delta_H}{\Delta_M}$.

So only the bank $n = 1$ most efficient in screening offers contracts for H projects that embed screening $C_{H,1}(R_{H,1}, B_{H,1}) \succ C_{H,n'}(R_{H,n'}, B_{H,n'})$ for $n' \in \{2, ..., N\}$. Note that bank screening while entrepreneurs choose the M project is not an equilibrium contract: since entrepreneurs adversely select their loan contract, they could always turn to a non-screening bank where they would still undertake the M project without supporting the cost of screening.

In the absence of screening, banks do not know the type e of the entrepreneur, and banks can only ensure that there are no incentives to undertake the L project. Under assumption 2, when a bank does not screen but still avoids moral hazard associated with gambling, the entrepreneur undertakes project M. Without screening, all banks have the same technology, so their PC is binding, and banks make zero profits:

$$p_M R_{M,n} B_{M,n} \quad \geq \quad r B_{M,n} \ . \tag{10}$$

And the interest rate is given by:[9]

$$R_{M,n} = \frac{r}{p_M} \quad \forall \quad n \in \{1, ..., N\} \ . \tag{11}$$

Competition on the extensive margin. The adverse selection process, whereby the design of a loan contract affects the risk-taking behaviour of entrepreneurs that pick either H or M projects, can be internalized by banks ex ante.

Due to heterogeneous screening costs, a rent is extracted by the most efficient bank $n = 1$, serving low-cost-of-effort entrepreneurs with project H. When no screening is undertaken, the equilibrium is such that banks make zero profit on entrepreneurs with a higher cost of effort that undertake project M. Thus the bank $n = 1$ most efficient

[9]If assumption 2 does not hold, banks that offer contracts C_M would finance mostly M projects but also some unscreened H projects. The expected probability of repayment in the absence of screening would then be given by $p_H > E(p) > p_M$ and $R_M = \frac{r}{E(p)}$. $E(p)$ would depend on the threshold type $\bar{e}_1$, above which entrepreneurs pick a contract from a bank that does not screen. Better lending conditions for H projects by screening banks would attract more customers into contract $C_{H,1}$, and this would reduce $E(p)$ for funding offered by non-screening banks. So R_M would increase, further increasing $\bar{e}_1$ with $R_M\left(\underset{+}{\bar{e}_1}(\underset{+}{R_M})\right)$. The intuition of the paper would not change.

at screening has an incentive to increase the market share of project H where it makes profits.

However, market shares of contracts financing H or M projects are not determined by bank, but by the adverse selection process of entrepreneurs. Thus bank $n = 1$ can increase its market share only by making loan contracts supporting H projects *relatively* more attractive by indirectly competing with itself. This is what I call self-competition.

Bank $n = 1$ can increase its market share $\bar{e}_1 - e_{min}$ by offering more attractive loan contracts with screening to the marginal entrepreneur since $\bar{e}_1 \left(\underset{-}{R_H} \right)$. A new threshold entrepreneur $\bar{e}_2$, defined as in equation (3), is indifferent between the two types of bank contracts, with or without screening. For entrepreneurs in $[\bar{e}_1, \bar{e}_2]$, bank $n = 1$ increases its profits by limiting its rent extraction until its marginal profit reaches zero $\Pi_1 (\bar{e}_2) = 0$. For entrepreneur $\bar{e}_2$, the interest rate $R_{H,1}$ similar to equation (9) is pinned down by s_1 instead of s_2.

Proposition 1. *Under assumptions 1 and 2, there exists a subgame-perfect Nash equilibrium so that the bank $n = 1$ most efficient at screening offers contracts $C_{H,1} (R_{H,1}, B_{H,1})$ to entrepreneurs with low costs of effort $e \leq \bar{e}_2$ that undertake projects H, while loan contracts without screening $C_{M,n} (R_{M,n}, B_{M,n})$ are offered by any bank $n \in \{1, ..., N\}$ with $N > 2$ and attract entrepreneurs with high costs of effort $e > \bar{e}_2$ undertaking M project.*

Proof. Given in the text. □

An illustration of the equilibrium contracts of this proposition is given by the plain line of Figure 2.[10]

The key assumption for the results is to have banks' rent-extraction ability on effort-making entrepreneurs increase with the efficiency of screening technologies. Banking

[10] In Figure 2, the loan volume offered to entrepreneurs just below and just above $\bar{e}_2$ is characterized by a break. This characteristic results from the fact that for those borrowers it becomes much more costly to provide effort: either they limit their own project size to limit their cost of effort if the solution is interior, or the bank limits their loan size to preserve incentives as the benefit of diverting increases with the loan size. One could argue that it is easier to ensure the success of a project of a reasonable scale, while excessively large projects are harder to manage and are less likely to succeed. However, to the left of $\bar{e}_2$, as entrepreneur's costs of effort decrease, the loan volume offered for H projects becomes much larger than the one offered to non-effort-making entrepreneurs undertaking the M project.

competition is a key ingredient as it determines the market structure and thus the importance of the presence of less efficient banks. If banks extract no rent, for instance because of costless entry or free screening technologies, then the impact of banks' relative efficiency vanishes. Then only the absolute efficiency of the screening bank matters.

Corollary 1. *In the absence of rent extraction, banking heterogeneity has no real effect.*

Proof. See the appendix. □

A notable by-product of banking heterogeneity is that it generates heterogeneity in the lending cycle. The cyclical pattern of aggregate bank lending depends on the strength of moral hazard for the different types of borrowers, as this affects the evolution of market shares of banks on the extensive margin. At the bank level, when varying economic conditions affect entrepreneurs with high effort costs, aggregate lending by banks that do not screen is procyclical. Conversely, the bank most efficient at screening is characterized by countercyclical aggregate lending. This is consistent with Beck et al. (2015), who show that banks that invest in soft information acquisition to reduce moral hazard have less volatile lending practices during economic downturns.

At the loan level, when economic downturns are characterized by increased moral hazard, loans for M projects with larger probabilities of default are more cyclical than those for H projects. This is consistent with empirical evidence suggesting that lower credit quality is associated with a larger volatility of credit (Berger and Udell, 2004; Dell'Ariccia et al., 2008).

Corollary 2. *Bank screening heterogeneity generates heterogeneity in the lending cycle.*

Proof. See the appendix. □

3.4 Specific case: duopoly

In the specific duopoly case where $N = 2$, the most efficient bank $n = 1$ faces an additional layer of self-competition. The direct self-competition effect is such that the marginal entrepreneur is incentivized to undertake the H project. The indirect self-competition effect is such that the marginal entrepreneur is discouraged from undertaking the M project.

Instead of competing for lending to entrepreneurs undertaking the M project, the most efficient bank $n = 1$ engages in a strategic reduction of competition across market segments. It seeks to offer loan contracts for M projects, where it does not make profits, only if it does not reduce its profitability.[11] The efficient bank $n = 1$ leaves some degree of market power to the less efficient bank $n = 2$ by deciding *not* to compete to finance M projects. As a result, the less efficient bank $n = 2$ extracts a rent with an offered interest rate $R_{M,2} > \frac{r}{p_M}$ above the competitive rate. With a larger $R_{M,2}$, fewer entrepreneurs choose the risky project M since $\frac{\partial \bar{e}_2}{\partial R_{M,2}} > 0$ and the new cut-off entrepreneur indifferent between projects H and M is now $\bar{e}_3$. The pool of screened customers serviced by bank $n = 1$ is now larger on $[e_{min}, \bar{e}_3]$.

The bank $n = 1$ most efficient in screening cannot deviate by providing an ϵ amount of loans to non-effort-making entrepreneurs undertaking project M. The sequencing of events of Figure 1 is such that banks play first and commit to the general terms of the loan contract. In addition, the screening costs are perfectly observable, so banks can form correct expectations about each other's subsequent behaviour.

The less efficient bank $n = 2$ maximizes its aggregate profit:[12]

$$R_{M,2} = \underset{R_{M,2} > r/p_M}{\texttt{argmax}} \int_{\bar{e}_3(R_{M,2})}^{e^{max}} \Pi_{M,2} \mathrm{d}F(e) = \left(e^{max} - \bar{e}_3\left(R_{M,2}\right)\right)\left(p_M R_{M,2} - r\right) B_{M,2}\left(R_{M,2}\right). \quad (12)$$

To pin down the equilibrium interest rate, the less efficient bank $n = 2$ faces the following trade-off: by using its market power, it extracts more rent per unit of loan (second term), but the size of each individual loan decreases (third term), and the number of customers choosing a risky project also decreases (first term) as contracts offered to entrepreneurs undertaking project M become relatively less attractive.

[11] All that is required for this additional effect is for the rent-extraction ability of banks to depend on the number of banks active on the market segment. The analysis would also apply to an oligopoly if $\frac{\partial R_M}{\partial N} < 0$.

[12] With a merger, in the absence of entry, the new entity would jointly maximize the profits of its two branches by internalizing the adverse selection effect: the equilibrium degree of monopoly power would be obtained by an equality of banks' marginal profits above or below the cut-off entrepreneur.

Proposition 2. *When the banking sector is concentrated ($N = 2$), under assumptions 1 and 2 there exists a fully separating subgame-perfect Nash equilibrium so that the bank $n = 1$ most efficient in screening offers contracts $C_{H,1}(R_{H,1}, B_{H,1})$ to entrepreneurs with a low cost of effort $e \leq \bar{e}_3$ that undertake project H, while the less efficient bank $n = 2$ does not screen and is the only one offering a loan contract $C_{M,2}(R_{M,2}, B_{M,2})$ that attracts entrepreneurs with a high cost of effort $e > \bar{e}_3$ undertaking project M.*

Proof. See the appendix. □

The equilibrium contracts when the banking sector is concentrated with $N = 2$ are illustrated by the dash-dotted line of Figure 2.

3.5 Optimal contract

A social planner internalizes the competition among banks across lending segments and the adverse selection by entrepreneurs, so self-competition disappears. The optimum corresponds to the set of loan contracts $\mathbb{C}_o$ a planner offers to entrepreneurs to maximize welfare W, namely total net production, while still facing the friction due to the unobserved cost of effort:

$$\underset{\bar{e}_o}{\text{Max}} \quad W = \int_{e_{min}}^{\bar{e}_o} \left((p_H\alpha - r)B_{H,o} - e\frac{B_{H,o}^2}{2} - s \right) \text{dF(e)} + \int_{\bar{e}_o}^{e_{max}} ((p_M\alpha - r)B_{M,o}) \,\text{dF(e)} \tag{13}$$

s.t. $B_{H,o}$ and $B_{M,o}$ from equations (5) and (7).

The first-order condition is given by:

$$(p_H\alpha - r)B_{H,o}(\bar{e}_o) - \bar{e}_o\frac{B_{H,o}(\bar{e}_o)^2}{2} - s = (p_M\alpha - r)B_{M,o} \ . \tag{14}$$

The cut-off entrepreneur $\bar{e}_o$ adds as much social value by undertaking the H or M project. This describes the optimum as the social surplus for project H decreases in e:

$$\frac{\partial W}{\partial e} = -\left((p_H\alpha - r) - eB_{H,o}\right) p_H \frac{\frac{\partial R_{H,o}}{\partial e}e + (\alpha - R_{H,o})}{e^2} - \frac{B_{H,o}^2}{2} < 0 \ , \tag{15}$$

where the last equation is indeed negative if $\frac{\partial R_{H,o}}{\partial e} > 0$ under assumption 1.

Assuming an interior solution, the combination of equations (3), (5) and (7) together with equation (14) yields:

$$(p_H R_{H,o} - r)\,\frac{p_H(\alpha - R_{H,o})}{\bar{e}_o} - (p_M R_{M,o} - r)\,\frac{2\Delta_M(\alpha - R_{M,o})}{b} = s\ . \tag{16}$$

From assumption 1, the marginal NPV of project H is larger than for project M, so the central planner wishes to maximize leverage of the entrepreneurs $[e_{min}; \bar{e}_o]$ undertaking project H. The interest rate $R_{H,o}$ is given by a zero-profits condition:

$$(p_H R_{H,o} - r)\, B_{H,o} = s\ . \tag{17}$$

$R_{H,o}$ is of the form given by equation(9), so it is easy to see that $\frac{\partial R_{H,o}}{\partial e} > 0$ holds. Equation (17) together with equation (16) pin down the value for $R_{M,o}$ given by:

$$(p_M R_{M,o} - r)\,\frac{2\Delta_M(\alpha - R_{M,o})}{b} = 0\ . \tag{18}$$

From assumption 1, one has $\alpha > R_{M,o}$, else M projects would not be financed despite the fact that they have a positive NPV. So it must be that $R_{M,o} = r/p_M$, and, as a result, the banking sector breaks even.

Proposition 3. *Under assumptions 1 and 2, the optimum is such that entrepreneurs $e < \bar{e}_o$ undertake H projects with the contract $C_{H,o}(B_{H,o}, R_{H,o})$ offered by the screening bank $n = 1$, while entrepreneurs $e > \bar{e}_o$ undertake M projects with the contract $C_{M,o}(B_{M,o}, R_{M,o})$ without screening.*

Proof. Given in the text. □

The dashed line of Figure 2 displays a representation of the optimal set of contracts.

4 Banking heterogeneity and banking risks

4.1 Distortions associated with banking heterogeneity

Entrepreneurs borrowing and moral hazard. First, from proposition 1, the presence of banking heterogeneity allows the most efficient bank to extract a rent over the fair pricing of the loan contract with a limit-pricing behaviour, while the optimum of proposition 3 is such that no rent is extracted. For entrepreneurs on $[e^{min}, \bar{e}_1]$ undertaking H projects, the bank has to reduce loan size ($\frac{\partial B_{H,1}}{\partial R_{H,1}} < 0$), either to adjust to the loan demand of the entrepreneur, or to avoid moral hazard if the EC or IC constraints are binding.

Second, from proposition 2, when the banking sector is in a situation of duopoly, banking heterogeneity creates a markup for loans to entrepreneurs undertaking M projects ($\frac{\partial B_{M,2}}{\partial R_{M,2}} < 0$). When a comparative advantage arises for one bank, market segmentation induces a strategic reduction of competition across other market segments.

Proposition 4. *On the intensive margin, heterogeneity in bank screening efficiency reallocates resources away from entrepreneurs compared with the optimum, towards screening banks, but also towards non-screening banks if the banking sector is concentrated ($N = 2$).*

Proof. Given in the text. □

However, this reallocation of resources from entrepreneurs to banks is not uniform across entrepreneurs. From proposition 1, entrepreneurs with high abilities are subject to stronger competition between banks: skilled entrepreneurs do not require a high compensation for effort, so the return of screening by banks is higher, and bank competition is stronger. Entrepreneurs with better abilities are less affected by the rent extraction associated with banking heterogeneity. In addition, the existence of self-competition is such that banks' rent-extraction ability is reduced for entrepreneurs with moderate skills; those entrepreneurs are close to being indifferent between switching to an M project without effort provision, and a larger rent extraction by the screening bank would decrease its market share.

Corollary 3. *Banking heterogeneity is most harmful to effort-making entrepreneurs in the middle of the distribution of effort costs around $\bar{e}_1$.*

Proof. Given in the text. □

Figure 3 displays the intensity of rent extraction over the distribution of entrepreneurs. The darker the area, the larger the lending premium over the fair lending rate offered at the optimum.

Effort decision and adverse selection. If market segmentation is not taken into account by the banking sector, the marginal entrepreneur is given by $\bar{e}_1$. Banking heterogeneity is associated with rent extraction by the most efficient bank $n = 1$, so providing effort becomes relatively less attractive, and some entrepreneurs change their risk-taking behaviour. Banking heterogeneity is associated with too little effort making ($\bar{e}_1 < \bar{e}_0$).

If market segmentation is taken into account via direct self-competition, the most efficient bank $n = 1$ competes more to increase its pool of customers. This restores optimality in the effort incentives of entrepreneurs by making sure that the screening bank extracts no rent from the marginal entrepreneur ($\bar{e}_2 = \bar{e}_o$).

In a concentrated banking sector ($n = 2$), indirect self-competition leads to a strategic reduction of competition on the market segment for entrepreneurs undertaking M projects. Some entrepreneurs with relatively low effort abilities may nonetheless undertake effort and pick the H project. Banking heterogeneity induces excessive effort making ($\bar{e}_3 > \bar{e}_0$).

Proposition 5. *On the extensive margin, compared with the optimum, heterogeneity in bank screening efficiency:*

(i) induces too few entrepreneurs $[e_{min}, \bar{e}_1]$ to undertake effort in the absence of self-competition with $\bar{e}_1 < \bar{e}_0$;

(ii) restores optimal effort decisions by entrepreneurs $[e_{min}, \bar{e}_2]$ in the presence of direct self-competition ($N > 2$) with $\bar{e}_2 = \bar{e}_0$;

(iii) induces too many entrepreneurs $[e_{min}, \bar{e}_3]$ to undertake effort in the presence of both direct and indirect self-competition ($N = 2$) with $\bar{e}_3 > \bar{e}_0$.

Proof. Given in the text. □

Figure 4 displays the effort choices of entrepreneurs as banking heterogeneity (Δ_s) increases on the horizontal axis. The dark grey area represents entrepreneurs that always undertake M projects without effort provision, while the light grey area represents entrepreneurs that always undertake H projects with effort. The hatched areas correspond to entrepreneurs that can change their project choice depending on the competition setup. The entrepreneur type indifferent between providing effort or not at the optimum is represented by the dashed-dotted line.

4.2 Non-performing loans as a metric for banking risks

The economy-wide share of non-performing loans (NPL) combines the intensive and extensive margins and writes, for a market segment cut-off $\bar{e}$:

$$\begin{aligned} NPL &= \frac{(1-p_H)\int_{e_{min}}^{\bar{e}} B_H \mathtt{d}F(e) + (1-p_M)\int_{\bar{e}}^{e_{max}} B_M \mathtt{d}F(e)}{\int_{e_{min}}^{\bar{e}} B_H \mathtt{d}F(e) + \int_{\bar{e}}^{e_{max}} B_M \mathtt{d}F(e)} \\ &= (1-p_H)\frac{1+\frac{1-p_M}{1-p_H}\frac{\int_{\bar{e}}^{e_{max}} B_M \mathtt{d}F(e)}{\int_{e_{min}}^{\bar{e}} B_H \mathtt{d}F(e)}}{1+\frac{\int_{\bar{e}}^{e_{max}} B_M \mathtt{d}F(e)}{\int_{e_{min}}^{\bar{e}} B_H \mathtt{d}F(e)}} . \end{aligned} \tag{19}$$

Denote $L = \frac{\int_{\bar{e}}^{e_{max}} B_M \mathtt{d}F(e)}{\int_{e_{min}}^{\bar{e}} B_H \mathtt{d}F(e)}$ the relative lending volumes for each project type. I have:

$$\frac{\partial NPL}{\partial L} = \frac{\Delta_H}{(1+L)^2} > 0 . \tag{20}$$

So the evolution of the ratio of non-performing loans depends on the evolution of the relative lending volumes L.

Corollary 4. *Productivity is procyclical, while the ratio of non-performing loans is countercyclical.*

Proof. See the appendix. □

Given propositions 4 and 5, the corollary follows:

Corollary 5. *Compared with the optimum NPL_o, the market equilibrium generates:*

(i) a larger share of NPL in the absence of self-competition;

(ii) a larger share of NPL in the presence of direct self-competition ($N > 2$);

(iii) a larger share of NPL via the intensive margin, but a lower share of NPL via the extensive margin in the presence of both direct and indirect self-competition ($N = 2$).

Proof. See the appendix. □

4.3 Banking efficiency improves

The screening ability of the less efficient bank $n = 2$ improves. A decrease in s_2 holding s_1 constant increases net production. The rent-extraction ability of the most efficient bank decreases as its comparative advantage proxied by the distance $\Delta_s = s_2 - s_1$ decreases. So the intensive margin of proposition 4 is such that fewer resources are taken away from effort-making entrepreneurs. The overall share of NPL decreases:

$$\frac{\partial NPL}{\partial s_2} = \frac{\partial NPL}{\partial L}\frac{\partial L}{\partial s_2} > 0 \ .$$

From equation (20), the first term is positive, and the second term is given by:

$$\frac{\partial L}{\partial s_2} = \frac{1}{\left(\int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e)\right)^2}\left(\underbrace{-\int_{e^{min}}^{\bar{e}} \frac{\partial B_H}{\partial s_2}\mathrm{d}F(e)\int_{\bar{e}}^{e^{max}} B_M \mathrm{d}F(e)}_{\text{intensive margin}} \right.$$
$$\left. \underbrace{-\frac{\partial \bar{e}}{\partial s_2}\left(B_M(\bar{e})\int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e) + B_H(\bar{e})\int_{\bar{e}}^{e^{max}} B_M \mathrm{d}F(e)\right)}_{\text{extensive margin}}\right) > 0 \ ,$$

where $\frac{\partial B_{H,1}}{\partial s_2} = \frac{\partial B_{H,1}}{\partial R_{H,1}}\frac{\partial R_{H,1}}{\partial s_2}$ is negative from equation (9). Improvement in an inactive screening technology can have real effects by limiting the dominant position of the most efficient bank $n = 1$. This increases welfare by reallocating resources to less risky borrowers undertaking the H project.

In the absence of self-competition, an additional effect arises on the extensive margin: from proposition 5, less rent extraction encourages effort making by the marginal entrepreneur ($\frac{\partial \bar{e}_1}{\partial s_2} < 0$ while $\frac{\partial \bar{e}_2}{\partial s_2} = 0$ and $\frac{\partial \bar{e}_3}{\partial s_2} = 0$), such that the decrease in NPL would be stronger.

The screening ability of the most efficient bank $n = 1$ improves. A decrease in s_1 holding s_2 constant increases net production as fewer resources are used by bank $n = 1$ in the screening process. The overall impact on NPL is given by:

$$\frac{\partial NPL}{\partial s_1} = \frac{\partial NPL}{\partial L}\frac{\partial L}{\partial s_1} .$$

From equation (20), the first term is positive, and the second term is given by:

$$\begin{aligned} \frac{\partial L}{\partial s_1} = \frac{1}{\left(\int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e)\right)^2} \Bigg(& \underbrace{\int_{\bar{e}}^{e^{max}} \frac{\partial B_M}{\partial s_1} \mathrm{d}F(e) \int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e)}_{\text{intensive margin}} \\ & \underbrace{- \frac{\partial \bar{e}}{\partial s_1} \left(B_M(\bar{e}) \int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e) + B_H(\bar{e}) \int_{\bar{e}}^{e^{max}} B_M \mathrm{d}F(e) \right)}_{\text{extensive margin}} \Bigg) . \end{aligned}$$

The effect on NPL of a lower screening cost s_1 now depends on bank competition.

In the absence of self-competition, the screening efficiency of the most efficient bank $n = 1$ is paradoxically irrelevant to the determination of the equilibrium contracts. As its screening technology improves, bank $n = 1$ makes more profits. But contracts are determined only by the screening ability of the competitor s_2 via limit pricing. So bank screening efficiency would have no direct real effects ($\frac{\partial B_{M,n}}{\partial s_1} = \frac{\partial \bar{e}_1}{\partial s_1} = 0$).

In the presence of direct self-competition ($N > 2$), the most efficient bank $n = 1$ makes loan contracts more attractive to the marginal entrepreneur ($\frac{\partial \bar{e}_2}{\partial s_1} < 0$). So effort making increases on the extensive margin, which decreases NPL, and the intensive margin plays no role ($\frac{\partial B_{M,n}}{\partial s_1} = 0$).

In the presence of direct and indirect self-competition ($N = 2$), the less efficient bank $n = 2$ also manages to extract a rent from non-effort-making entrepreneurs, but the better the most efficient bank, the smaller the share of non-effort making entrepreneur, the lower the rent-extraction ability of the less efficient bank as entrepreneurs are now less captive and more likely to switch banks. Then the loan volume granted to riskier borrowers undertaking the M project increases ($\frac{\partial B_{M,2}}{\partial s_1} < 0$). So the intensive effect is such that the share of NPL increases. However, the extensive margin is still decreasing the share of NPL ($\frac{\partial \bar{e}_3}{\partial s_1} < 0$), so the overall effect is ambiguous.

Proposition 6. *(i) Increasing banking efficiency is associated with increased production efficiency.*

(ii) Increasing the screening efficiency of the less efficient bank $n = 2$ increases financial stability.

(iii) When increasing the screening efficiency of the most efficient bank $n = 1$: (a) in the absence of self-competition, there is no impact on financial stability; (b) in the presence of direct self-competition ($N > 2$), it increases financial stability; (c) in the presence of direct and indirect self-competition ($N = 2$), the effect on financial stability is ambiguous.

Proof. Given in the text. □

4.4 Heterogeneity in banking efficiency decreases

I now turn to the case where the proxy for bank screening heterogeneity $\Delta_s = s_2 - s_1$ varies in a mean-preserving spread fashion, keeping average screening efficiency across the first two banks constant.[13] The overall effect of banking heterogeneity on NPL is:

$$\frac{\partial L}{\partial \Delta_s} = \frac{1}{\left(\int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e)\right)^2} \Bigg(\underbrace{\int_{\bar{e}}^{e^{max}} \frac{\partial B_M}{\partial \Delta_s} \mathrm{d}F(e) \int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e) - \int_{e^{min}}^{\bar{e}} \frac{\partial B_H}{\partial \Delta_s} \mathrm{d}F(e) \int_{\bar{e}}^{e^{max}} B_M \mathrm{d}F(e)}_{\text{intensive margin}}$$

$$\underbrace{- \frac{\partial \bar{e}}{\partial \Delta_s} \left(B_M(\bar{e}) \int_{e^{min}}^{\bar{e}} B_H \mathrm{d}F(e) + B_H(\bar{e}) \int_{\bar{e}}^{e^{max}} B_M \mathrm{d}F(e) \right)}_{\text{extensive margin}} \Bigg) .$$

The impact of a decrease in banking heterogeneity (lower Δ_c) on financial stability depends on the bank competition set-up.

In the absence of self-competition, the share of NPL decreases towards the optimum. On the intensive margin, the rent-extraction ability of the most efficient bank $n = 1$ is reduced (proposition 4). The loan volume offered to effort-making entrepreneurs undertaking the H project increases ($\frac{\partial B_{H,1}}{\partial \Delta_c} < 0$). This leads to a decrease in the relative weight given to riskier lending. As a result, on the extensive margin, more entrepreneurs are incentivized to make effort and undertake project H that is less risky ($\frac{\partial \bar{e}_1}{\partial \Delta_c} < 0$).

In the presence of direct self-competition, the effect on the share of NPL is ambiguous. The effect on the intensive margin is similar, but now the effect on the extensive margin is as follows: a mean-preserving decrease of banking heterogeneity that reduces rent-extraction abilities makes the most efficient bank $n = 1$ less able to compete and attract the marginal entrepreneur. Fewer entrepreneurs are incentivized to make effort ($\frac{\partial \bar{e}_2}{\partial \Delta_c} > 0$), which contributes to increasing the share of NPL.

In the presence of direct and indirect self-competition ($N = 2$), the effect on the share of NPL is also ambiguous. On top of the opposite forces from the intensive and extensive margins presented above, an additional effect mitigates the adverse impact on the share of NPL. As the marginal entrepreneur $\bar{e}_3$ decreases, it increases the rent-extraction ability of

[13] A mean-preserving change in the relative screening technologies is equivalent to $\frac{\partial s_1}{\partial s_2} = -1$. Note that screening costs for the banks $n \in \{3, ..., N\}$ with $N \geq 3$ play no role in this simple set-up. If the choice of productive projects had more dimensions than just two (H and M), then the market for loans could be partitioned into more than two segments. Then screening-cost differences among more than two banks would matter.

the less efficient bank $n = 2$, so the interest rate $R_{M,2}$ for the riskier M project increases, which in turn increases the marginal entrepreneur $\bar{e}_3$.

In the presence of market segmentation, self-competition arises such that a decrease in banking heterogeneity is associated with an increase in the share of NPL on the extensive margin. But it is also associated with a decrease in the share of NPL on the intensive margin. Different magnitudes of the two margins lead to a trade-off between productive efficiency, banking efficiency and banking risks.

Proposition 7. *Less heterogeneity in bank screening efficiency can generate a trade-off between production efficiency and financial stability:*

(i) production efficiency improves as rent extraction decreases;

(ii) in the absence of self-competition, financial stability improves towards the optimum;

(iii) in the presence of self-competition, the effect on financial stability is ambiguous and depends on the relative strength of the intensive and extensive margins.

Proof. Given in the text. □

Figure 5 provides a numerical illustration of the trade-off between production efficiency and financial stability in the presence of self-competition. For low banking heterogeneity (to the left on the figure), the intensive margin effect dominates, and a reduction of banking heterogeneity is associated with a lower share of NPL that converges to the optimum. For higher banking heterogeneity (to the right on the figure), the extensive margin effect dominates, and a reduction of banking heterogeneity is associated with an increase in the share of NPL compared with the optimum.

5 Discussions

5.1 Scope for regulation

One can think of a macroprudential authority with a mandate to avoid excessive lending to riskier borrowers or maintain sector-wide banking losses at the optimum. The regulator would then try to improve on the market equilibrium by imposing a macroprudential levy $\{\tau_H, \tau_M\}$ on bank lending to bring the share of aggregate NPL closer to the optimum, subject to a break-even constraint:

$$\min_{\{\tau_H,\tau_M\}} |NPL - NPL_o| \tag{21}$$
$$\text{s.t.} \int_{\bar{e}}^{e^{max}} \tau_M B_M \mathrm{d}F(e) = \int_{e^{min}}^{\bar{e}} \tau_H B_H \mathrm{d}F(e) \ .$$

From corollary 5, it is straightforward to see that in the absence of self-competition or with direct self-competition ($N > 2$), a higher tax on banks with more lending to entrepreneurs that undertake the M project associated with a larger probability of default can restore optimality of the share of NPL. The balance sheet constraint of banks lending to entrepreneurs undertaking M projects is now:

$$(p_M R_M - r - \tau_M)\, B_M = 0 \Leftrightarrow R_M = \frac{r + \tau_M}{p_M} \ . \tag{22}$$

A larger $\tau_M > 0$ yields a larger R_M that lowers B_M and increases $\bar{e}$ from equation (3). This in turn reduces the share of NPL from equation (19). Therefore, provided that the market equilibrium is such that $NPL > NPL_o$, a tax on loans to riskier borrowers can restore optimality in the share of NPL. In order to satisfy the budget constraint with $\tau_M > 0$, one must subsidize lending to less risky entrepreneurs who are undertaking the H project $\tau_H < 0$. This is equivalent to subsidizing the screening bank $n = 1$ offering loans to entrepreneurs undertaking the H project. However, from proposition 1, the heterogeneous tax will increase $R_{H,1}$ as it is pinned down by the funding cost of bank $n = 2$ only. The subsidy will lead to higher profit margins for the screening bank without any improvement in the lending conditions for high-ability entrepreneurs undertaking the H project.

With a macroprudential authority, bank self-competition prevents a subsidy to the screening bank from being rebated to the customers undertaking H projects. However, a fiscal authority could both restore financial stability and improve the allocation of credit by implementing a tax scheme on both lenders and borrowers to ensure better financing conditions for the entrepreneurs with better abilities. By subsidizing directly the borrowers that undertake the H project, especially those with intermediate abilities (corollary 3), the fiscal authority successfully reduces entrepreneurs' effective cost of borrowing despite the rent extraction by the screening bank $n = 1$. Still, this is possible only if the fiscal authority itself has access to the information on borrowers' type or on the precise terms of the loan contract.

Proposition 8. *Let's consider an authority averse to economy-wide excessive loan losses* $NPL > NPL_o$.

(i) A macroprudential authority can successfully restore optimality with respect to the financial stability criterion, but it harms production efficiency.

(ii) A fiscal authority can successfully restore optimality with respect to the financial stability criterion and also restore production efficiency.

Proof. Given in the text. □

5.2 Costly entry in a concentrated banking sector

So far entry was not allowed. We focus here on the most interesting case where entry is allowed in a concentrated banking sector ($N = 2$) so that it reshapes banking competition.

Entry of a bank with screening cost s_η could occur if it pays a cost η.[14] Entry by a bank with a better screening technology $s_\eta < s_1$ occurs if $\Pi_{H,\eta}(R_{H,\eta}, s_\eta) > \eta$. In this case, the pricing of loans to effort-making borrowers is lower $R_{H,\eta}(s_1) < R_{H,1}(s_2)$ and now pinned down by the lowest interest rate the incumbent bank $n = 1$ can offer, given its screening cost s_1. As the incumbent bank $n = 1$ loses its market share for H projects, it now competes to finance M projects; the indirect self-competition effect disappears, and the interest rate R_M is also lower.

The impact of entry on the share of NPL now depends on the relative variation of the competition intensity in all market segments. If the new interest rate $R_{H,\eta} = R_{H,1} - \epsilon$ for ϵ is small enough, which arises for large enough entry costs η, then lending conditions

[14] Barriers to entry or informational rents are usually put forward as reasons that prevent the entry of competitors and limit technological changes (Dell'Ariccia and Marquez, 2004; Sengupta, 2007).

ease relatively more for M projects than for H projects; the cut-off entrepreneur $\bar{e}_3$ is such that adverse selection tightens and more entrepreneurs stop providing effort. The overall market share for financing M projects increases relatively more, and, as a result, the share of NPL increases.

When the banking sector is concentrated, the threat of entry of a less efficient bank with screening costs $s_\eta > s_1$ can also reduce the rent-extraction ability of loans offered for M projects that do not require screening. The interest rate $R_{M,2}$ now has to satisfy the no-competition condition of equation (12) together with the no-entry condition $p_M B_{M,2} R_{M,2} - r B_{M,2} - \eta = 0$. If this last condition is binding, the threat of entry decreases the market power of banks when financing M projects. Adverse selection unambiguously tightens: as $R_{M,2}$ decreases, more entrepreneurs stop providing effort, and $\bar{e}_3$ gets lower. The bank-wide share of NPL increases.

Proposition 9. *When the banking sector is concentrated ($N = 2$):*

(i) costly bank entry improves allocation efficiency by limiting bank rent-extraction abilities;

(ii) entry by a bank more efficient at screening $s_\eta < s_1$ deteriorates financial stability if the fixed costs of entry η is large enough;

(iii) the threat of entry by a bank less efficient at screening $s_\eta > s_1$ deteriorates financial stability.

Proof. Given in the text. □

Boyd and De Nicolo (2005) show that more banking competition unambiguously decreases bank risk taking. Proposition 9 instead shows that bank entry, although successful at limiting the negative externality associated with banking heterogeneity, can be associated with more financial instability if the banking sector is concentrated, depending on the relative evolution of competition across different market segments. Thus, increasing banking competition by decreasing barriers to entry in a concentrated banking sector may not be undertaken by a policy-maker more concerned with financial stability.

5.3 Relevance to alternative situations

Although presented in the context of banks more or less efficient at screening, the model can provide an insight into alternative types of heterogeneity that generate the same intensive and extensive margin effects.

Heterogeneity due to foreign bank entry. Developing economies usually face more heterogeneity, given less efficient local banks and international competitors entering the market with better technologies. As the technological gap increases following foreign entry, the more efficient foreign banks have a less risky loan portfolio compared with other banks, credit to the private sector tends to be lower on the intensive margin, and domestic banks seek new market niches on the extensive margin (Bonin and Abel, 2000; Detragiache et al., 2008).

However, this can also reflect the case of more developed economies such as the United States during the 1980s and 1990s. The deregulation on bank branches across states generated the entry of larger banks seeking to expand their market share by challenging local banks with dominant positions. This led to a better allocation of loans with higher loan volumes to more productive projects, but at the same time it generated a significant rise in bankruptcy rates (Dick and Lehnert, 2010).

Heterogeneity between banks and non-banks. The rise of shadow banking is a by-product of new financial innovations (Gorton and Metrick, 2011) allowing specialized financial intermediaries to compete with traditional banks. As the informational advantage of the traditional banking sector fades away, too many projects are financed via shadow banks without adequate incentives; the extensive margin calls for an increase in the share of NPL. But higher competition forces the traditional banking sector to lower its rent extraction. The intensive margin calls for a reduction of the share of NPL. A more efficient shadow banking sector can indirectly improve the allocation of credit by the traditional banks (Loutskina, 2011), but it may also lead to an increase in the share of NPL as the market share of the shadow banks increases (Luck and Schempp, 2014).[15]

[15] However, shadow banks affect financial stability in many other ways (e.g., Adrian et al., 2002).

Merging to become more efficient. A merger would alter the heterogeneity in the banking sector. Berger et al. (1998) found that a merged entity tends to focus on lending to higher-quality borrowers and reduce its provision of loans to smaller borrowers. On the extensive margin, lending to small business is picked up by the non-consolidated entities, but those banks are likely to provide less favourable loan terms in the intensive margin with less efficient lending relationships (Berger et al., 2001).

Relationship versus transaction lending. One can consider that the informed bank with better screening technologies provides relationship lending. Other fund providers behave more like financial markets. They simply make transactions to channel funds to the borrowers. As competition increases, the rent-extraction ability of the relationship lending bank decreases (Schenone, 2010), and the profits of the bank are redirected towards borrowers that enjoy higher leverage. Thus the institution providing relationship lending is less able to attract the marginal entrepreneur, and its market share decreases (Boot and Thakor, 2000).

Heterogeneity driven by regulations. The choice of banks between different risk management approaches within the Basel regulation can lead to some degree of bank specialization and loans-market segmentation. Repullo and Suarez (2004) show that high-risk-profile firms will prefer to borrow from banks that adjust less their loan profile to individual risks, that is banks that adopt the standardized approach as opposed to the internal ratings-based approach that allows banks to fine-tune loan risk metrics. When banks are given the choice between two different approaches for the rating of loan risks, a competitive advantage arises for the banks that invest in the technology that allows them to fine-tune risk metrics (Hakenes and Schnabel, 2011). As a result, banks that adjust their loan profile less to individual risks specialize in riskier lending. This may in turn generate higher aggregate risks.

6 Conclusion

This paper focuses on the financial stability impact of market segmentation generated by banks' heterogeneous access to borrowers' information. Production efficiency and financial stability do not necessarily go hand in hand, and a trade-off can emerge: banking heterogeneity unambiguously decreases welfare away from the optimum, but banking heterogeneity can distort incentives towards a lower economy-wide share of non-performing loans closer to the optimum. When banking heterogeneity increases, limit-pricing competition on the intensive margin increases the economy-wide share of non-performing loans, but direct or indirect self-competition on the extensive margin decreases the economy-wide share of non-performing loans. This self-competition effect creates a tension between the intensive and extensive margins.

The results have several implications. First, entry improves allocation efficiency, but the impact of bank competition on banking risks is non-linear due to market segmentation. Second, ex ante screening generates a specific type of market segmentation: entrepreneurs with low costs of providing effort are more likely to be screened, while those entrepreneurs are usually less likely to be monitored. Third, a macroprudential authority averse to excessive loan losses can successfully restore financial stability by applying a bank-specific tax scheme targeted towards riskier lenders, but this harms production efficiency. Fourth, the entrepreneurs that suffer most from increased banking heterogeneity are those with intermediate effort costs, so policies aiming at restoring production efficiency should target specific classes of borrowers.

One main caveat is that the simple framework presented here does not model the liability side of banks, so financial stability is merely reflected by the overall share of non-performing loans. Banking stability should also be analyzed against the capitalization of the banking sector that can withstand an increase in non-performing loans.

Going forward, more theoretical and empirical work is needed to understand the interaction between screening and monitoring for different classes of borrowers, the role of bank competition in the presence of self-selection of heterogeneous bank loan contracts, or the impact of mergers and acquisition of banks with heterogeneous technologies.

References

Adrian, T., A. B. Ashcraft, and N. Cetorelli (2002) “Shadow Bank Monitoring,” in A. Berger, P. Molyneux, and J. Wilson eds. *The Oxford Handbook of Banking*: Oxford University Press. 15

Beck, T., H. Degryse, R. De Haas, and N. Van Horen (2015) “When Arm’s Length Is Too Far. Relationship Banking over the Credit Cycle,” CEPR working paper. 1, 3.3

Berger, A. N., L. G. Goldberg, and L. J. White (2001) “The Effects of Dynamic Changes in Bank Competition on the Supply of Small Business Credit,” *European Finance Review*, Vol. 5, pp. 115–139. 5.3

Berger, A. N., A. Saunders, J. M. Scalise, and G. F. Udell (1998) “The Effects of Bank Mergers and Acquisitions on Small Business Lending,” *Journal of Financial Economics*, Vol. 50, No. 2, pp. 187–229. 5.3

Berger, A. and G. Udell (2004) “The Institutional Memory Hypothesis and the Procyclicality of Bank Lending Behavior,” *Journal of Financial Intermediation*, Vol. 13, No. 4, pp. 458–495. 1, 3.3

Bester, H. (1985) “Screening vs. Rationing in Credit Markets with Imperfect Information,” *American Economic Review*, Vol. 74, No. 4, pp. 850–855. 2

Bonin, J. and I. Abel (2000) “Retail Banking in Hungary: A Foreign Affair?” William Davidson Institute Working Paper No. 356. 5.3

Boot, A. and A. Thakor (2000) “Can Relationship Banking Survive Competition?” *Journal of Finance*, Vol. 55, No. 2, pp. 679–713. 1, 5, 5.3

Boyd, J. and G. De Nicolo (2005) “The Theory of Bank Risk Taking and Competition Revisited,” *Journal of Finance*, Vol. 60, No. 3, pp. 1329–1343. 1, 5.2

Cetorelli, N. and P. Peretto (2000) “Oligopoly Banking and Capital Accumulation,” Federal Reserve Bank of Chicago, mimeo. 2

Dell’Ariccia, G., E. Detragiache, and R. Rajan (2008) “The Real Effect of Banking Crises,” *Journal of Financial Intermediation*, Vol. 17, No. 1. 1, 3.3

Dell’Ariccia, G. and R. Marquez (2004) “Information and Bank Credit Allocation,” *Journal of Financial Economics*, Vol. 72, No. 1, pp. 185–214. 14

Detragiache, E., T. Tressel, and P. Gupta (2008) “Foreign Banks in Poor Countries: Theory and Evidence,” *Journal of Finance*, Vol. 63, No. 5, pp. 2123–2160. 5.3

Diamond, D. (1984) “Financial Intermediation and Delegated Monitoring,” *Review of Economic Studies*, Vol. 51, pp. 393–414. 2.2

Dick, A. and A. Lehnert (2010) “Personal Bankruptcy and Credit Market Competition,” *Journal of Finance*, Vol. 65, No. 2, pp. 655–686. 1, 5.3

Fama, E. (1985) “What’s Different About Banks?” *Journal of Monetary Economics*, Vol. 15, pp. 29–39. 2.2

Gorton, G. and A. Metrick (2011) "Regulating the Shadow Banking System," *Brookings Papers on Economic Activity*, Vol. 2, pp. 261–312. 5.3

Hakenes, H. and I. Schnabel (2011) "Bank size and risk-taking under Basel II," *Journal of Banking and Finance*, Vol. 35, No. 6, pp. 1436–1449. 1, 5.3

Hauswald, R. and R. Marquez (2003) "Information Technology and Financial Services Competition," *Review of Financial Studies*, Vol. 16, No. 3, pp. 921–948. 3

——— (2006) "Competition and Strategic Information Acquisition in Credit Markets," *Review of Financial Studies*, Vol. 19, No. 3, pp. 967–1000. 4

Holmstrom, B. and J. Tirole (1997) "Financial intermediation, loanable funds and the real sector," *Quarterly Journal of Economics*, Vol. 112, No. 3, pp. 663–691. 2.1

Hughes, J. and L. Mester (2009) "Efficiency in Banking: Theory, Practice, and Evidence," in A. Berger, P. Molyneux, and O. Wilson eds. *The Oxford Handbook of Banking*, Oxford. 7

Loutskina, E. (2011) "The role of securitization in bank lending and funding management," *Journal of Financial Economics*, Vol. 100, pp. 663–684. 5.3

Luck, S. and P. Schempp (2014) "Banks, Shadow Banking and Fragility," ECB working paper 1726. 5.3

Martinez-Miera, D. and R. Repullo (2015) "Search for Yield," CEPR Discussion Paper No. 10830. 5

Petersen, M. (2004) "Information: Hard and Soft," Kellogg School of Management, mimeo. 1

Petersen, M. A. and R. G. Rajan (1995) "The effect of credit market competition on lending relationship," *Quarterly Journal of Economics*, Vol. 110, No. 2, pp. 407–443. 5

Repullo, R. and J. Suarez (2004) "Loan pricing under Basel capital requirements," *Journal of Financial Intermediation*, Vol. 13, No. 4, pp. 496–521. 5.3

Schenone, C. (2010) "Lending Relationships and Information Rents: Do Banks Exploit Their Information Advantages?" *Review of Financial Studies*, Vol. 23, No. 3, pp. 1149–1199. 3, 5.3

Sengupta, R. (2007) "Foreign entry and bank competition," *Journal of Financial Economics*, Vol. 84, No. 2, pp. 502–528. 14

Stein, J. (2002) "Information production and capital allocation: Decentralized vs. hierarchical firms," *Journal of Finance*, Vol. 57, pp. 1891–1921. 1

A Proofs

Proof of Corollary 1. In the case of proposition 1, the interest rate schedule is given by:

$$\begin{cases} R_{H,1}\left(s_2\right) \quad \text{on} \quad e \in [e_{min}, \bar{e}_1] \\ R_{H,1}\left(s_1, s_2\right) \quad \text{on} \quad e \in [\bar{e}_1, \bar{e}_2] \\ R_{M,n} \quad \text{with} \quad n \in \{1, ..., N\} \quad \text{on} \quad e \in [\bar{e}_2, e_{max}] \end{cases} . \tag{23}$$

If, instead, the bank $n = 1$ most efficient at screening does not extract a rent, the interest rate schedule is given by:

$$\begin{cases} R_{H,1}\left(s_1\right) \quad \text{on} \quad e \in [e_{min}, \bar{e}_1] \\ R_{M,n} \quad \text{with} \quad n \in \{1, ..., N\} \quad \text{on} \quad e \in [\bar{e}_1, e_{max}] \end{cases} . \tag{24}$$

One can see that heterogeneity in the screening costs s_1 and s_2 matters only when bank $n = 1$ can extract a rent from its customers. □

Proof of Corollary 2. In good times, the return of providing effort increases, so p_H gets larger (similarly, if the cost e is scaled down). $\frac{\partial R_{H,1}}{\partial p_H} < 0$ is equivalent, after rearranging, to the sign of $p_H\alpha(3p_H\alpha - 2r) - r^2 > 0$, which holds under assumption 1. So it implies $\frac{\partial B_{H,1}}{\partial p_H} > 0$ from equation (5) on the intensive margin, and from equation (3) together with assumption 1 one has $\frac{\partial(\bar{e} - e_{min})}{\partial p_H} > 0$ on the extensive margin. So aggregate lending to entrepreneurs undertaking the H project by the bank $n = 1$ most efficient at screening is procyclical. Lending without screening to entrepreneurs undertaking the M project is countercyclical since the extensive margin is countercyclical $\frac{\partial(e_{max} - \bar{e})}{\partial p_H} > 0$ and the intensive margin is inelastic $\frac{\partial B_{M,n}}{\partial p_H} = 0$.

In good times, moral hazard b for entrepreneurs not undertaking effort decreases. From equations (7) and (5), $\frac{\partial B_{M,n}}{\partial b} < 0$ and $\frac{\partial B_{H,1}}{\partial b} = 0$, so individual loans offered to entrepreneurs undertaking the M project more likely to default are more cyclical. From equation (3), $\frac{\partial(e_{max} - \bar{e})}{\partial b} < 0$. Thus the intensive and extensive margins for banks financing M projects decreases, and aggregate lending for M projects is procyclical. Lending with screening for H projects is countercyclical since $\frac{\partial(\bar{e}-e_{max})}{\partial b} < 0$.

$\square$

Proof of Proposition 2. The participation constraint of entrepreneurs undertaking the M project is $\alpha > R_{M,2}$ which gives an upper bound for the rent-extraction ability. From equation (7), $\frac{\partial B_{M,2}}{\partial R_{M,2}} < 0$, so the marginal profits of entrepreneurs undertaking an M project is given by:

$$\frac{\partial \pi_M}{\partial R_{M,2}} = p_M \left((\alpha - R_{M,2}) \frac{\partial B_{M,2}}{\partial R_{M,2}} - B_{M,2} \right) < 0 \ .$$

From an equation similar to (3) defining $\bar{e}_3 \left(R_{H,1}, B_{H,1}, R_{M,2}, B_{M,2}\right)$, I have $\frac{\partial \bar{e}_3}{\partial R_{M,2}} > 0$. As the market power of the bank funding M projects increases, incentives for entrepreneurs to undertake the M projects decreases, and some turn to project H.

The less efficient bank $n = 2$ chooses its interest rate $R_{M,2}$ so that:

$$R_{M,2} = \underset{R_{M,2} \in]\frac{r}{p_M};\alpha[}{\texttt{argmax}} \ (e^{max} - \bar{e}_3) \left(p_M R_{M,2} - r\right) B_{M,2} \quad (25)$$

subject to $B_{M,2}$ given by equation (7).

The first derivative with respect to $R_{M,2}$ is given by:

$$-\frac{\partial \bar{e}_3}{\partial R_{M,2}} \left(p_M R_{M,2} - r\right) B_{M,2} + \left(e^{max} - \bar{e}_3\right) \left(p_M B_{M,2} + (p_M R_{M,2} - r) \frac{\partial B_{M,2}}{\partial R_{M,2}} \right) \ . \quad (26)$$

For $R_{M,2} = \frac{r}{p_M}$, equation (26) becomes:

$$\left(e^{max} - \bar{e}_3\right) p_M B_{M,2} > 0 \ .$$

For $R_{M,2} \to \alpha > \frac{r}{p_M}$, the first term of equation (26) is negative, and the second term is also negative and given by:

$$\underset{R_{M,2} \to \alpha}{lim} \left(p_M B_{M,2} + (p_M R_{M,2} - r) \frac{\partial B_{M,2}}{\partial R_{M,2}} \right) = \frac{2\Delta_M}{b} \left(r - 2 p_M \alpha\right) < 0 \ .$$

Provided the variables are continuous, there exists at least one interior solution with a positive markup such that $\frac{r}{p_M} < R_{M,2} < \alpha$.

There is no incentive to deviate from the equilibrium. For the less efficient bank, its technology $s_2 > s_1$ is less efficient, so it would never screen. For the efficient bank, recall that the bank decides to screen first, and then it learns the type of the entrepreneur it faces. Upon screening, the bank has to use the information it acquired to adjust its loan. Similarly, entrepreneurs have no incentive to deviate. An entrepreneur has no incentive to stop providing efforts if the bank learned his type, since the contract has a built-in effort-compatible condition. Second, the entrepreneur does not provide effort if he is not screened by the bank, as it violates assumption 2.

□

Proof of Corollary 4. Similar to the share of NPL, aggregate productivity writes:

$$Pvity = \frac{\alpha p_H \int_{e^{min}}^{\bar{e}} B_H \mathtt{d}F(e) + \alpha p_M \int_{\bar{e}}^{e^{max}} B_M \mathtt{d}F(e)}{\int_{e^{min}}^{\bar{e}} B_H \mathtt{d}F(e) + \int_{\bar{e}}^{e^{max}} B_M \mathtt{d}F(e)} = \alpha p_H \frac{1 + \frac{p_M}{p_H} L}{1 + L} \ . \quad (27)$$

During upturns, the return of providing effort increases, so p_H gets larger:

$$\frac{\partial Pvity}{\partial p_H} = \alpha \frac{1 + K - \Delta_H \frac{\partial L}{\partial p_H}}{(1+L)^2} > 0 \quad \text{and} \quad \frac{\partial NPL}{\partial p_H} = \frac{-1 - L + \Delta_H \frac{\partial L}{\partial p_H}}{(1+L)^2} < 0 \ .$$

The demand for loans for H projects increases ($\frac{\partial B_H}{\partial p_H} > 0$ and $\frac{\partial R_H}{\partial p_H} < 0$), and entrepreneurs find it more profitable to make effort, so $\bar{e}$ increases. So $\frac{\partial L}{\partial p_H} < 0$. □

Proof of Corollary 5. From propositions 4 and 5 and the definition of the share of NPL given in (19):

(i) in the absence of self-competition: $\bar{e}_1 < \bar{e}_o$, $B_{H,1} < B_{H,o}$, $B_{M,n} = B_{M,o}$ so $NPL > NPL_o$;

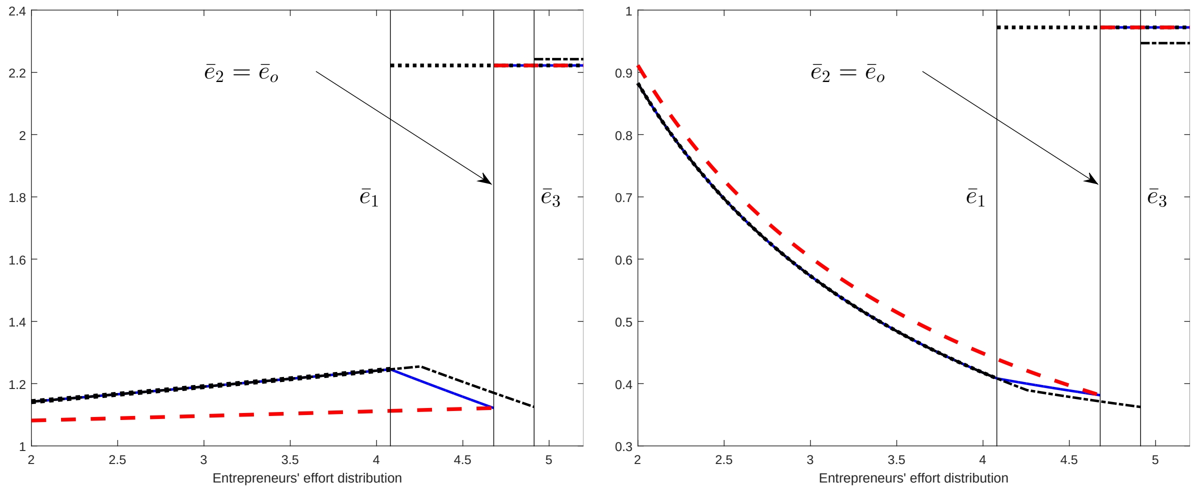
(ii) in the presence of direct self-competition ($N > 2$): $\bar{e}_2 = \bar{e}_o$, $B_{H,1} < B_{H,o}$, $B_{M,n} = B_{M,o}$ so $NPL > NPL_o$;

(iii) in the presence of both direct and indirect self-competition ($N = 2$): $\bar{e}_3 > \bar{e}_o$, $B_{H,1} < B_{H,o}$, $B_{M,2} < B_{M,o}$ so NPL >or< NPL_o. □

B Simulations

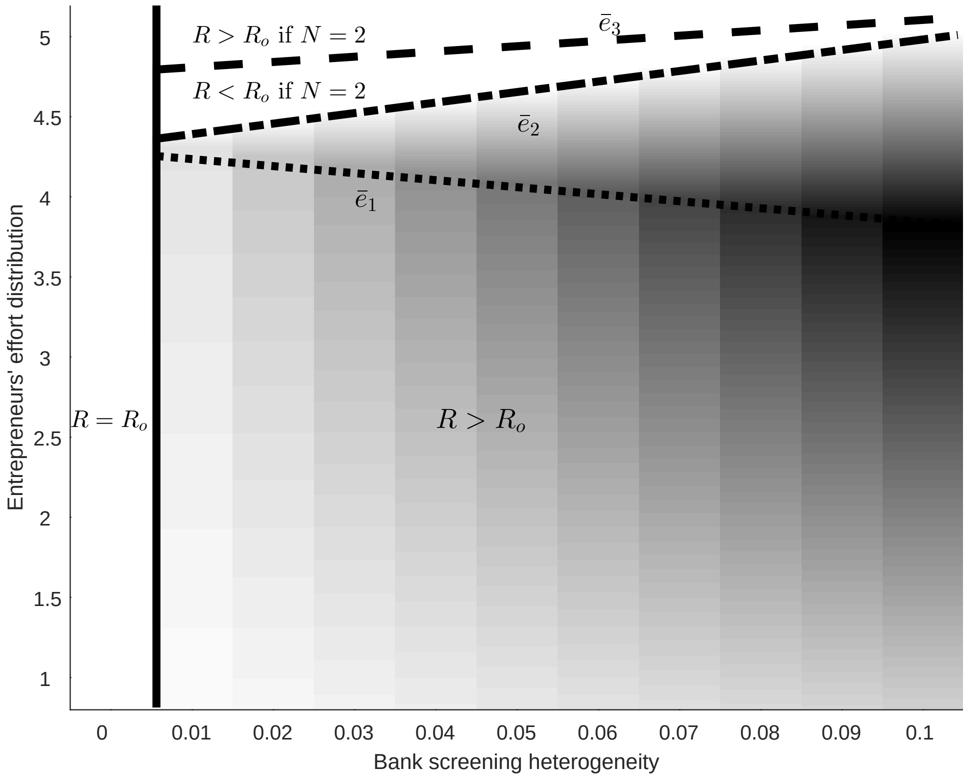
Figure 2: The set of bank loan contracts with banking heterogeneity

(a) Lending rate **(b)** Loan volume



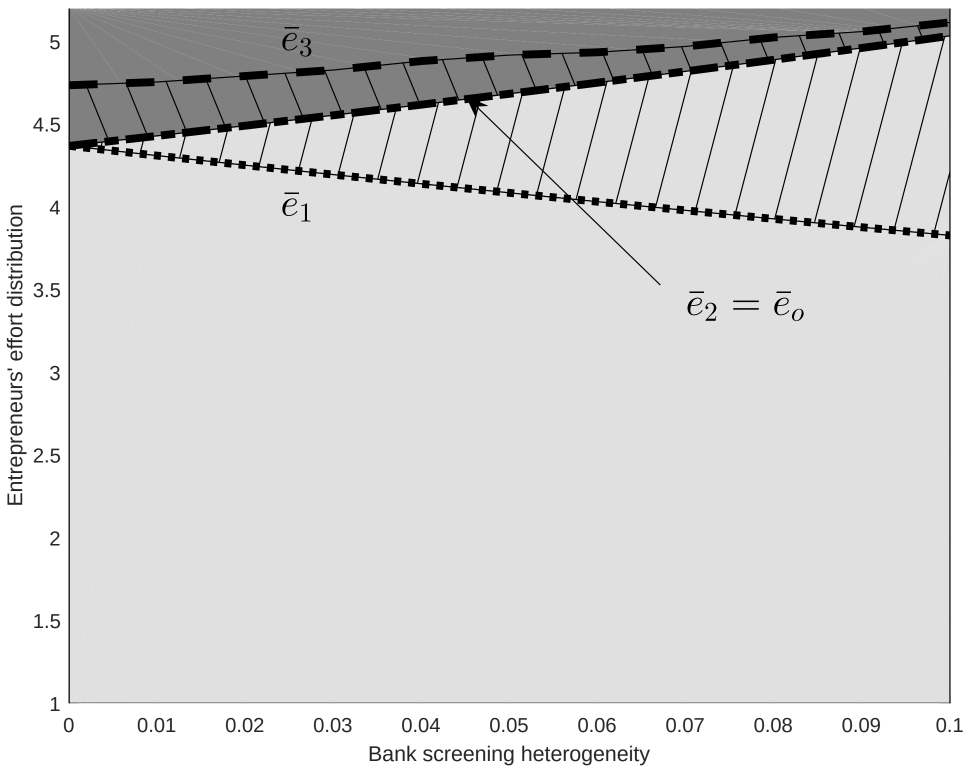
This figure displays the set of bank loan contracts for the market equilibrium in the absence of self-competition (dotted), the market equilibrium with self-competition if $N > 2$ (plain) or $N = 2$ (dash-dotted), and the optimum (dashed) for each entrepreneur type e on the horizontal axis. The numerical example is for $p_H = 0.95$, $p_M = 0.45$, $p_L = 0.2$, $\alpha = 3$, $b = 0.4$, $s_1 = 0.025$, $s_2 = 0.075$, $r = 1$, $e^{min} = 0.8$ (the chart is censured to the left for sake of readability), $e^{max} = 5.2$ that satisfy assumptions 1 and 2. Market segmentation is given by the vertical bars. To the left of a vertical bar, screening is undertaken by the most efficient bank and entrepreneurs choose project H, while to the right, no screening is undertaken and entrepreneurs find themselves better off choosing project M.

Figure 3: Bank rent extraction with banking heterogeneity



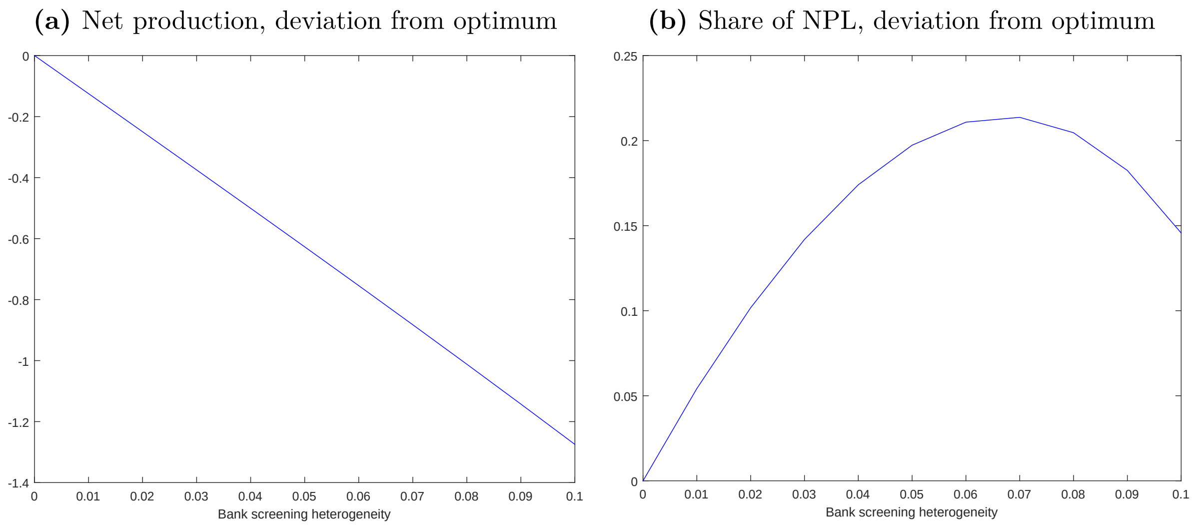
This figure displays the strength of the rent extraction of banks above the optimal interest rate as banking heterogeneity (Δ_s) increases on the horizontal axis. The numerical example is for $p_H = 0.95$, $p_M = 0.45$, $p_L = 0.2$, $\alpha = 3$, $b = 0.4$, $s_1 = 0.05 - \Delta_s/2$, $s_2 = 0.05 + \Delta_s/2$, $r = 1$, $e^{min} = 0.8$, $e^{max} = 5.2$ that satisfy assumptions 1 and 2. The darker the area the higher the premium over the fair lending rate R_o at the optimum. For each intensity of banking heterogeneity, the premium is largest along the dotted line representing $\bar{e}_1$. If the banking sector is concentrated ($N = 2$), then entrepreneurs with high costs of effort $e > \bar{e}_3$ above the dashed line also face a lending premium, while entrepreneurs with slightly lower effort costs $\bar{e}_2 < e < \bar{e}_3$ between the dashed and dashed-dotted line have a discount because they now switch lending categories.

Figure 4: Entrepreneurs' choice with banking heterogeneity



This figure displays the effort choices of entrepreneurs as banking heterogeneity (Δ_s) increases on the horizontal axis. The numerical example is for $p_H = 0.95$, $p_M = 0.45$, $p_L = 0.2$, $\alpha = 3$, $b = 0.4$, $s_1 = 0.05 - \Delta_s/2$, $s_2 = 0.05 + \Delta_s/2$, $r = 1$, $e^{min} = 0.8$, $e^{max} = 5.2$ that satisfy assumptions 1 and 2.
At the optimum, the dashed-dotted line represents the cut-off entrepreneurs $\bar{e}_o$ indifferent between providing effort or not. Henceforth, the dark grey area represents projects M, while the light grey area represents projects H. The hatched area corresponds to entrepreneurs that can change their project choice depending on the competition set-up when the market equilibrium deviates from the optimum.
If the banking sector is not too concentrated ($N > 2$), the non-hatched area in light grey corresponds to projects H financed by the bank most efficient at screening that makes profits via limit-pricing competition. The hatched area in light grey corresponds to project H financed by the bank most efficient at screening that makes profits via direct self-competition.
If the banking sector is concentrated ($N = 2$), the hatched area in dark grey corresponds to H projects financed by the bank most efficient at screening that makes profits via indirect self-competition. The non-hatched dark grey area corresponds to M projects financed by the bank less efficient at screening that makes profits due to the indirect self-competition effect.

Figure 5: Production efficiency versus banking risks



This figure displays the market equilibrium as banking heterogeneity (Δ_s) increases on the horizontal axis. The numerical example is for $p_H = 0.95$, $p_M = 0.45$, $p_L = 0.2$, $\alpha = 3$, $b = 0.4$, $s_1 = 0.05 - \Delta_s/2$, $s_2 = 0.05 + \Delta_s/2$, $r = 1$, $e^{min} = 0.8$, $e^{max} = 5.2$ that satisfy assumptions 1 and 2. The sub-figures represent the deviation of the production net of costs (effort, screening, funding) and the share of non-performing loans in percentage points compared with the social optimum.