Heterogeneous Banking Efficiency : Allocative Distortions and Lending Fluctuations

Thibaut Duprey

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

WORKING PAPER N° 2013 – 39

Heterogeneous Banking Efficiency : Allocative Distortions and Lending Fluctuations

Thibaut Duprey

JEL Codes : J12, D63, D31, D12

Keywords: Banking heterogeneity, Moral hazard, Adverse selection, Endogenous market segmentation, Allocation efficiency, Lending cycle

PARIS-JOURDAN SCIENCES ECONOMIQUES
48, BD JOURDAN – E.N.S. – 75014 PARIS
TÉL. : 33(0) 1 43 13 63 00 – FAX : 33 (0) 1 43 13 63 10
www.pse.ens.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE – ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES
ÉCOLE DES PONTS PARISTECH – ECOLE NORMALE SUPÉRIEURE – INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE

Heterogeneous Banking Efficiency : Allocative Distortions and Lending Fluctuations *

Duprey Thibaut†

*I am grateful to Xavier Ragot for guidance, and wish to thank Régis Breton, Zhili Cao, Nicolas Coeurdacier, Enisse Kharroubi, Guillaume Plantin, Annukka Ristiniemi, Daria Shakourzadeh for helpful comments on previous drafts. This paper also benefited from comments by seminar participants at the Paris School of Economics, the Banque de France, the 2013 AFSE-LAGV Congress in Aix-en-Provences as well as the 2013 EEA Congress in Gothenburg. I gratefully acknowledge financial support from the Banque de France. The opinion expressed herein may not reflect the position of the Banque de France. Any remaining errors are mine.

†Paris School of Economics and Banque de France. E-mail : thibaut.duprey@gmail.com.

Résumé

Cet article analyse les fluctuations et l'allocation du crédit en présence de banques hétérogènes dans leur capacité à gérer le risque de contrepartie pour leurs crédits aux entreprises : une distribution du crédit moins pro-cyclique n'assure pas une allocation efficace du crédit. Le mécanisme analysé est double. (a) La rente dont bénéficie la banque la plus efficace modifie les incitations à l'effort des entrepreneurs. (b) Les banques qui n'assurent pas de suivi de l'entrepreneur ont un volume de prêts agrégé moins sensible aux chocs de productivité. En conséquence : (i) la présence d'un système bancaire hétérogène décroit la productivité moyenne des entreprises car cela augmente la sélection adverse par les entrepreneurs et favorise l'extraction de rentes par les banques ; (ii) une banque ayant un volume de prêts moins cyclique signale une moindre efficacité de ses capacités de gestion du risque ; (iii) une distribution du crédit moins cyclique sur le plan agrégé peut ne pas être désirable si elle est associée à une distorsion des incitations et de l'allocation du crédit résultant d'un système bancaire hétérogène.

Mots clés : hétérogénéité bancaire, hazard moral, selection adverse, segmentation de marché endogène, allocation du crédit, cycle du crédit.

Codes JEL : G21, E30.

Abstract

This paper is a first attempt to connect the heterogeneity in bank efficiency with lending fluctuations and allocation efficiency : there is a trade-off between the two in the presence of heterogeneity in bank monitoring efficiency. The mechanism at hand is twofold. (a) First the rent extracted by the most efficient bank distorts incentives of entrepreneurs to undertake efforts. (b) Second banks specialising on contracts that do not include monitoring feature less cyclical fluctuations of aggregate lending. This has clear implications: (i) the presence of banking heterogeneity decreases firms' average productivity as it increases adverse selection by entrepreneurs as well as favours rent extractions by banks; (ii) an individual bank featuring a lower cyclicality signals a lower efficiency in its monitoring abilities; (iii) a heterogeneous banking system featuring a lower cyclicality of aggregate lending might not be desirable as it may come along with allocative and incentives distortions.

Keywords : banking heterogeneity, moral hazard, adverse selection, endogenous market segmentation, allocation efficiency, lending cycle.

JEL Classification : G21, E30.

Executive summary

This paper is a first attempt to connect heterogeneity in bank efficiency with lending fluctuations and allocative efficiency, and I show that there is a trade-off between the two when banks are heterogeneous. More precisely I introduce a heterogeneous wedge in the monitoring cost of banks and look at the consequences for entrepreneurs' effort choice, individual and aggregate credit allocation as well as the evolution of productivity in equilibrium. This type of heterogeneity across banks is akin to a negative externality because banks' monitoring decisions spill over all market segments –including those that do not necessarily require monitoring– through entrepreneurs' adverse selection of banks' loan contracts. Nevertheless both efficient and less efficient banks still exist in equilibrium, as less efficient banks, if they cannot be competitive when distinguishing high from medium effort-making entrepreneurs, still have a competitive technology to distinguish medium effort-making entrepreneurs from non-productive projects.

Thus, heterogeneity in bank lending fluctuations reflects the heterogeneity on the asset side with projects of variable profitability; in other words, heterogeneity in bank monitoring efficiency provides incentives for banks to specialise in lending to specific entrepreneurs. The differences in monitoring costs generate an endogenous market segmentation which allows the most efficient bank to extract a rent on a niche. On the one hand, this rent distorts credit allocation and incentives : some entrepreneurs find themselves better off reducing the intensity of their effort and thus choose loan contracts outside of the niche that do not include monitoring. On the other hand, banks that do not monitor entrepreneurs feature less cyclical fluctuations of aggregate lending: more entrepreneurs choose these less efficient banks in bad times as they do not want to make efforts (extensive margin), and these banks reduce less the loan size offered to entrepreneurs as lower effort is relatively less detrimental to incentives in bad times (intensive margin). Overall, increased banking heterogeneity reduces the fluctuations of the economy as increased rent extraction reduces the share and the size of the procyclical loan contracts granted to effort-making entrepreneurs.

This has clear implications:

1. First, when one observes some banks featuring a lower cyclicality of their aggregate lending, it may not be desirable to favour their development to obtain less cyclicality as it may signal a lower efficiency in their monitoring abilities.
2. Second, if some countries seemed to have been more resilient to the early phases of the

crisis by not cutting aggregate lending so much which resulted in smaller variations of output, it may be at the cost of a very heterogeneous banking sector associated with allocation inefficiencies.

3. Third, overall banking efficiency is not enough to ensure allocation efficiency, so that a more granular approach is required to make a sound assessment of a banking sector's efficiency, as more banking heterogeneity is detrimental to aggregate productivity.
4. Fourth, increasing banking concentration, i.e. giving them more market power on the main segment of the market, can reduce the rent extraction in the niche, but is detrimental to overall resources allocation.
5. Last, reforms implementing new regulations which would have an impact on monitoring or screening incentives (such as capital or liquidity regulation) should also be careful not to increase banking heterogeneity that would create possibilities of rent extractions and distort the adverse selection process by borrowers and as a result dampen lending fluctuations at the cost of increased allocation inefficiency.

1 Introduction

Structural reforms of the banking sector focus on defining the scope of allowed activities, both to increase the resilience and the stability of the banking system. But this approach does not tackle the issue of the heterogeneity of the banking sector within its boundaries. Before pushing forward reforms aiming at fostering the development of activities deemed as more stable, one should first investigate the source of their apparent stability and the associated consequence for allocation efficiency.

Indeed, lending fluctuations and productivity levels may not go hand in hand. The finance and growth literature mostly backs the idea that deeper financial markets (Levine, 2005; Rajan and Zingales, 1998; Wurgler, 2000) and a larger financial sector (King and Levine, 1993; Demirgüç-Kunt et al., 2008; Hartmann et al., 2007) would have a positive impact on capital allocation efficiency and growth. However, since the Great Recession, it becomes more widely accepted that there is a tradeoff with financial stability, whereby a more developed financial sector increases volatility (Kose et al., 2003; Levchenko et al., 2009), with crises usually being preceded by rapid out-of-trend growth in financial aggregates (Kaminsky and Reinhart, 1999; Alessi and Detken, 2009).

Some new evidences suggest that cross-sectional differences in the financial sector, especially banking heterogeneity, did influence the cyclical behaviour of the economy. Heterogeneity in bank efficiency seemed to help countries weather the first part of the crisis, that is to say the presence of a less efficient banking system limited the reduction in output growth (Giannone et al., 2011). Moreover, some specific banking activities such as investment banking are known to be more volatile than standard commercial banking (Adrian and Shin, 2010); likewise, banks with specific ownership characteristics feature different lending patterns, with governmental owned banks (Duprey, 2012; Bertay et al., 2012) or banks active only on a local segment of the market (Micco and Panizza, 2006) having a lending policy less responsive to macroeconomic shocks. So banking heterogeneity seems to be a source of fluctuations, and lending fluctuations across banks is itself heterogeneous. But this additional layer of (heterogeneous) cyclicality may not coincide with credit allocation efficiency by channelling too much funds either to excessively risky investments or to mismanaged firms not properly incentivised to behave well.

In this paper, I investigate the extent to which heterogeneity in bank efficiency matters for lending fluctuations at the aggregate and across banks, as well as the implied reallocation

of credit in the economy and the consequence for overall productivity; I show that increasing banking heterogeneity creates a trade-off between dampening the sensitivity of lending to productivity shocks and reaching an allocation of credit ensuring productive efficiency. To that extent, I develop a framework based on contract theory with three main ingredients, namely moral hazard, adverse selection and rent extraction, in order to study variations of banking heterogeneity on both the cross-sectional and time dimension of credit allocation. The wedge comes from the introduction of heterogeneity in monitoring abilities which thus creates a monitoring efficiency externality different across banks: introducing a wedge in monitoring efficiencies on the market that requires monitoring may spill over to markets which do not require monitoring, and thus modify bank loan contracts even when monitoring is not required. Note that the model reproduces the realistic feature of a countercyclical markup on the cost of loans, positive profits for banks due to rent extraction and limited competition, as well as procyclical individual loan contracts.

The model includes heterogeneity along three dimensions. First I consider several financial intermediaries (from the duopoly to the n-bank case) whose only difference is their position in the distribution of monitoring cost. Second I depart from the standard finance literature by considering more than one positive NVP project, so that I have two layers of moral hazard, corresponding respectively to the choice of effort intensity which is simplified to be binary, and the choice of shirking or not which is represented by a negative NPV project with private benefits. This heterogeneity in NPV's projects allows to create a phenomenon of niche; else, with a single positive NPV project, I would fall back in the traditional Darwinian process where the less efficient banking institutions never enter the market. Third I have a continuum of entrepreneurs with different effort abilities which introduces adverse selection. It makes sure that some entrepreneurs will always be less keen to pay the private effort cost so that they would prefer to decrease their effort intensities by not choosing the largest positive NVP project; they can do so by not revealing their type to the banks and avoiding monitoring.

The story goes as follows. Banks with larger monitoring costs are *de facto* less efficient than their counterparts but do not necessarily disappear in equilibrium. This wedge creates an endogenous market segmentation ; as the efficient bank has a comparative advantage in monitoring entrepreneurs in order to reduce moral hazard, competition will force less efficient competitors out of this market segment. It then allows the efficient bank to extract a rent over the fair interest repayment. As a result only entrepreneurs with a sufficiently small cost of effort

relative to the gains from a larger probability of success will self-select this contract inducing the maximum amount of effort. In other word, the most efficient bank finds itself in a niche from which it can extract a rent on high productivity entrepreneurs.

The remaining active firms, which would undertake the minimal amount of effort, will naturally turn to the banking sector that does not monitor actively entrepreneurs. If banks' market power is positive and bank entry has a significant impact on competition, that is to say the banking sector is not fully competitive, then the less efficient banks will be the only ones active on the segment of the market for low-effort entrepreneurs. Indeed the following rent seeking and adverse selection argument applies : the most efficient bank would be better off extracting the largest rent it can on effort-making entrepreneurs, so it should lock in its customers by making sure other segments of the market are not too attractive for them. The efficient bank can do so by preventing itself from entering other markets; else it would increase banking competition in other market segments which would result in more attractive contracts offered to entrepreneurs outside the niche; this would ultimately provide an incentive for entrepreneurs previously in the niche and willing to make the highest level of effort to choose now the minimum required level of effort, on which only a smaller or no rent can be extracted.

Thus I obtain a separating equilibrium, with the tail (least able) entrepreneurs choosing to undertake the minimum required level of effort and self-selecting the contracts designed by less efficient banks, while entrepreneurs with good effort abilities choose to make more effort and thus prefer to self-select banking contracts offered by the most efficient bank that allow them to reveal their type truthfully upon monitoring [1]. Henceforth, the heterogeneity in lending fluctuations reflects the heterogeneity in the productive sector, that is to say the composition of the asset side [2]; then heterogeneity in monitoring efficiency provides incentives for banks to specialise in lending to specific entrepreneurs. Note that both efficient and less efficient banks still exist in equilibrium, as less efficient banks, if they cannot be competitive when distinguishing high from medium effort-making entrepreneurs, still have a competitive technology to distinguish medium effort-making from negative NPV projects associated with private benefits to entrepreneurs, hence the need to have a richer NPV project structure.

The source of fluctuation in this model is threefold. First I have the standard source of

1. In other words, there is a sort of complementarity between effort provided by entrepreneurs and monitoring undertaken by the banker, i.e. both need to get involved to make sure the high yield project is undertaken. This is especially true for small and medium size firms, and could be understood as a sort of one period lending relationship.

2. I abstract from heterogeneity on the liability side.

cyclicality highlighted in Holmstrom and Tirole (1997) which is that the moral hazard itself is countercyclical, hence the benefit of deviating decreases when the economy expands. So whether or not I have banking heterogeneity, I will still obtain procyclical variations in the model.

Second the structure of the contract leads to heterogeneity in lending fluctuations. On the one hand, due to softer moral hazard considerations for entrepreneurs making less effort, individual loan amounts are less responsive to productivity shocks for less efficient bank contracts (intensive margin); while on the other hand, in periods of recession, adverse selection worsens so that more entrepreneurs decide to switch to the minimum effort level allowed, i.e. from efficient to less efficient bank loan contracts (extensive margin). Henceforth, in case of productivity shock, both the size of the loan and the number of customers selecting efficient bank contracts decrease, while for less efficient banks the size of the loan decreases less than for efficient bank contracts and the pool of entrepreneurs at the minimum effort level self-selecting less efficient bank contracts increases. So at the aggregate level, efficient bank lending is more responsive to productivity shocks –more procyclical– than bank lending by less efficient banks.

Third, the specific source of fluctuation added here is due to the bank heterogeneity in monitoring abilities which leads to adverse selection: endogenous decisions by competing banks to monitor or not, when monitoring abilities are known but heterogeneous across banks, create a dampening effect steaming from market segmentation. The more heterogeneity there is between bank monitoring technology, i.e. the less efficient the technology for some banks, the lower the aggregate cyclicality compared to the optimum. It can be understood as the additional rent extraction by the most efficient bank reduces the number of effort-making entrepreneurs, that is to say it reduces the share of the most pro-cyclical financial institution, and overall sensitivity to productivity changes is reduced.

Then the allocative distortion associated with banking heterogeneity is as follows. Because of the endogeneous switching decisions between several positive NPV projects, the productivity shock –homogeneous for all projects– can result into a heterogeneous reaction of bank loan contracts and modify the distribution of effort choices, creating a composition effect; thus the common productivity shock may not necessarily map one to one into aggregate observed productivity. The presence of heterogeneity in bank monitoring efficiency decreases average productivity as it increases adverse selection by entrepreneurs (extensive margin) as well as favours rent extractions by banks (intensive margin). Henceforth, the reallocation of loans towards entrepreneurs making less effort and the incentives distortions induced by the creation

of a niche lead to an unambiguous negative impact on output levels, which is here the only possible metric to analyse allocation efficiency. Last, if the market power of the less efficient banks was to increase, it could partly get the fluctuations and incentives right by preventing switches due to rent extraction differential across market segments, but it would take further resources away from the most able entrepreneurs willing to make effort.

Five main policy implications are straightforward from the model. First, when one observes some banks featuring a lower cyclicality of their aggregate lending, it may not be desirable to favour their development to obtain less cyclicality as it may signal a lower efficiency in their monitoring abilities. Second, if some countries seemed to have been more resilient to the early phases of the crisis by not cutting aggregate lending so much which resulted in smaller variations of output, it may be at the cost of a very heterogeneous banking sector associated with allocation inefficiencies. Third, overall banking efficiency is not enough to ensure allocation efficiency, so that a more granular approach is required to make a sound assessment of a banking sector's efficiency, as more banking heterogeneity is detrimental to aggregate productivity. Fourth, increasing banking concentration, i.e. giving them more market power on the main segment of the market, can reduce the rent extraction in the niche, but is detrimental to overall resources allocation. Last, reforms implementing new regulations which would have an impact on monitoring or screening incentives (such as capital or liquidity regulation) should also be careful not to increase banking heterogeneity that would create possibilities of rent extractions or induce adverse selection and as a result decrease lending fluctuations at the cost of larger allocation inefficiencies.

The next section provides the main stylized facts on banking heterogeneity and banking efficiency. Section 3 presents the heterogeneous banking model with moral hazard, adverse selection and rent seeking. First I derive the equilibrium for two benchmarks, the homogeneous banking case (section 4) and the social equilibrium (section 5). Then I study in section 6 the equilibrium with two heterogeneous banks before extending it in section 7 by considering several banks. Last, section 8 emphasizes the main sources of distortion associated with banking heterogeneity, both in terms of allocation and fluctuation. Section 9 provides numerical simulations and section 10 concludes.

2 Stylized facts

By banking heterogeneity, I mean heterogeneity in monitoring efficiency via differences in bank abilities to manage high yield loans or different costs of doing so. So as a proxy for cost efficiency in loan management, I use the variable Overhead costs from the bank specific database Bankscope [3] over unconsolidated balance sheet statements from 1995 to 2010, thus including 27540 banks in 178 countries. Then to be able to focus on the heterogeneity in banking efficiency, I use the standard deviation [4] of the Overhead over Asset ratio (in constant 2005 USD) within each country for each year; thus I actually plot 662 country year observations. A standard deviation close to zero would mean that the banking sector is more homogeneous in its cost efficiency. Instead of the raw correlation, the displayed coefficient for the slope of the regression line on each graph includes country and year fixed effect to remove unnecessary noise from the large panel and thus strenghten the stylised facts.

The novel set of stylised facts outlined here, which is the focus of the modelling, is the following:

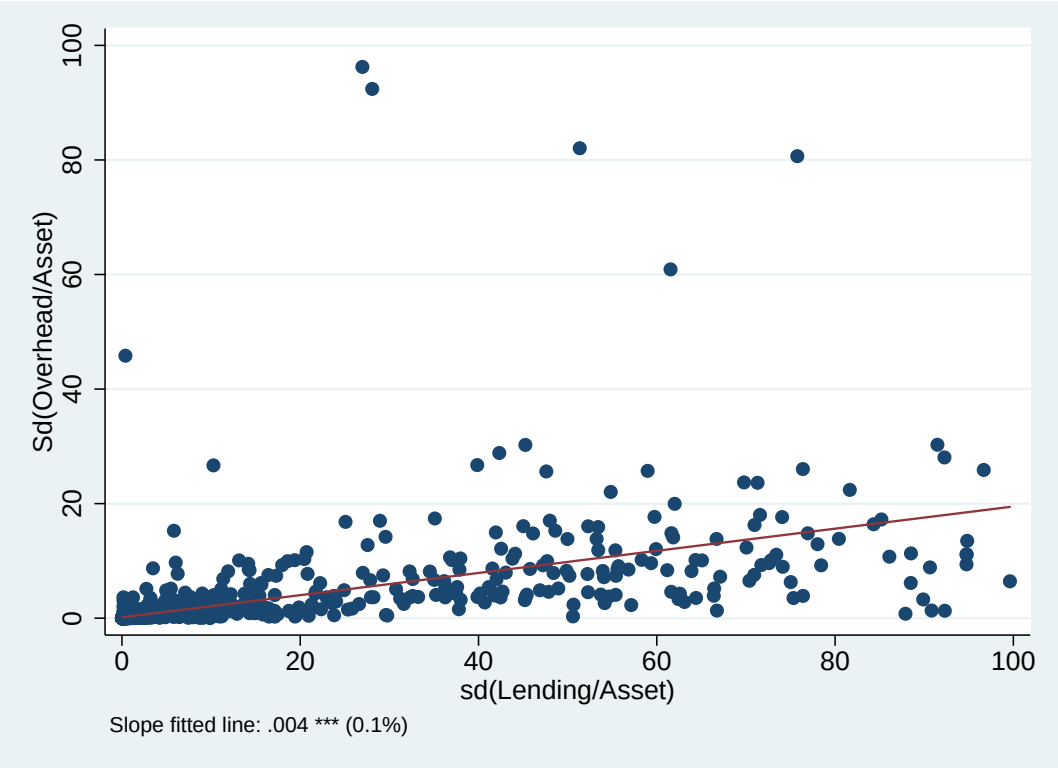
1. **bank lending decision are heterogeneous across banks**. Thus heterogeneous banking efficiency maps into heterogeneous lending decisions, which is proxied by larger standard deviation of the ratio of aggregate bank lending over total assets (see graph 1);

2. **heterogeneous banking efficiency is associated with allocation inefficiencies**. At the bank level, a more heterogeneous banking sector maps into lower lending aggregated (see graph 2). At the country level, a more heterogeneous banking sector is possibly detrimental to productivity : productivity is proxied by the output gap which results from an H-P filter with smoothing parameter 6.5 on USD 2005 constant GDP taken from the World Bank statistics (see graph 3). Thus with banking heterogeneity being associated with lower lending volumes and lower production intensities, a more heterogeneous banking tend to be also associated with lower GDP growth (see graph 4).

3. **heterogeneous banking efficiency is associated with lower lending fluctuations**. At the bank level, it has been outlined by Duprey (2012) and Bertay et al. (2012) in the specific case of public banks which are largely accepted to be detrimental to long term macroeconomic variables since the seminal paper by La Porta et al. (2002). Likewise

3. For a more precise description of the database, see Duprey and Lé (2012).
4. I exclude standard deviation of the ratio above 100 which can be considered as outliers. Also, I drop growth rates of assets above 100% which likely reflects mergers and acquisitions.

Figure 1: Heterogeneous banking efficiency maps into heterogeneous lending volatilities



national banks feature lower lending fluctuations than foreign owned banks that usually implement better banking practices (Micco and Panizza, 2006). At the country level, the work by Giannone et al. (2011) suggests that a less efficient banking sector was associated with a lower output drop in the early phase of the Great Recession.

3 Description of the model

The model I present here follows the lines of Holmstrom and Tirole (1997) (thereafter HT) although it departs from the original paper on several important points. First the incentive structure is expressed in the same way as Reichlin (2004). Then I focus on a distribution of effort cost –i.e. effort abilities– among the entrepreneurs rather than a distribution of collateral, in order to emphasize how the efficiency of the banking sector translates into productive efficiency [5]. But the main deviation from the original HT paper is that I allow for two projects, with two effort intensities, to have a positive –but different– net present value.

5. Note that with a variable investment scale, which is the case here, firms with different levels of assets would use the same optimal policy scaled by their asset; hence assuming an identical level of collateral across all entrepreneurs is not restrictive here and only makes the exposition easier.

Figure 2: Heterogeneous banking efficiency maps into lower bank lending

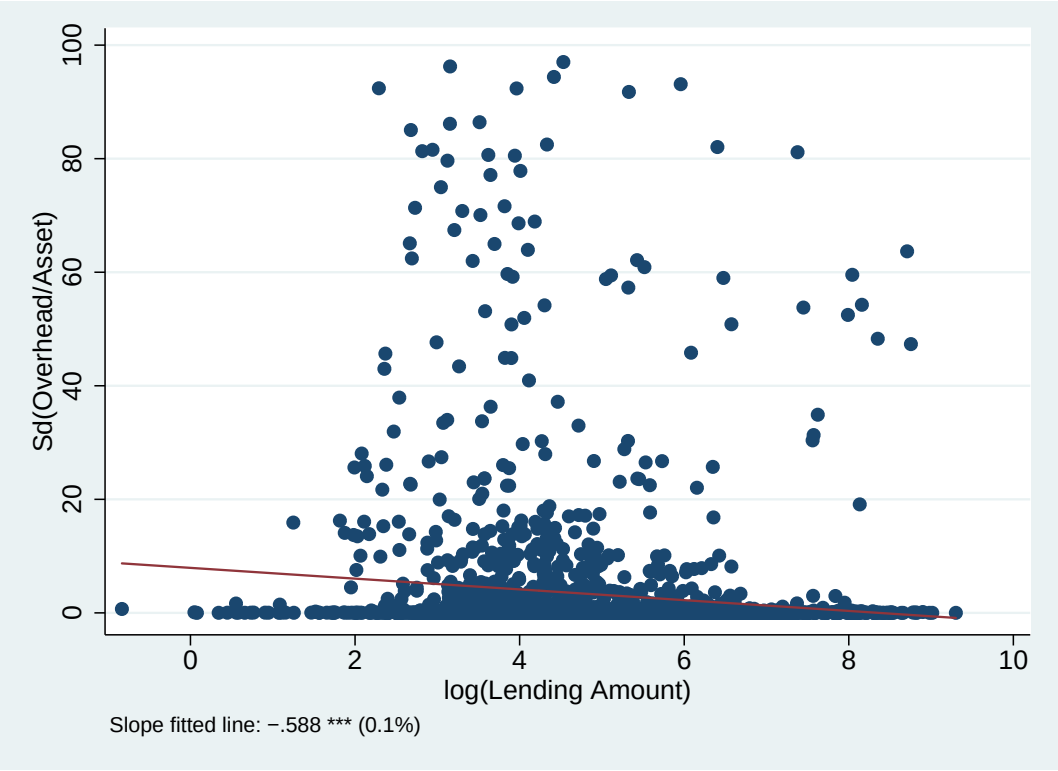


Figure 3: Heterogeneous banking efficiency maps into lower productivity

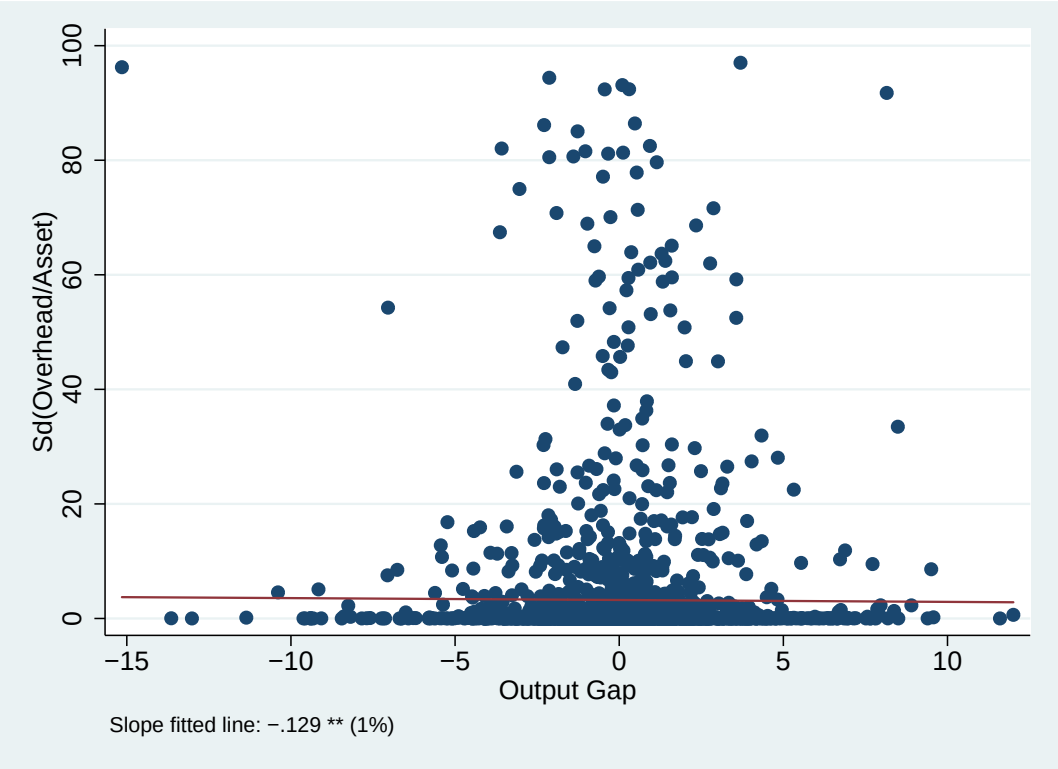
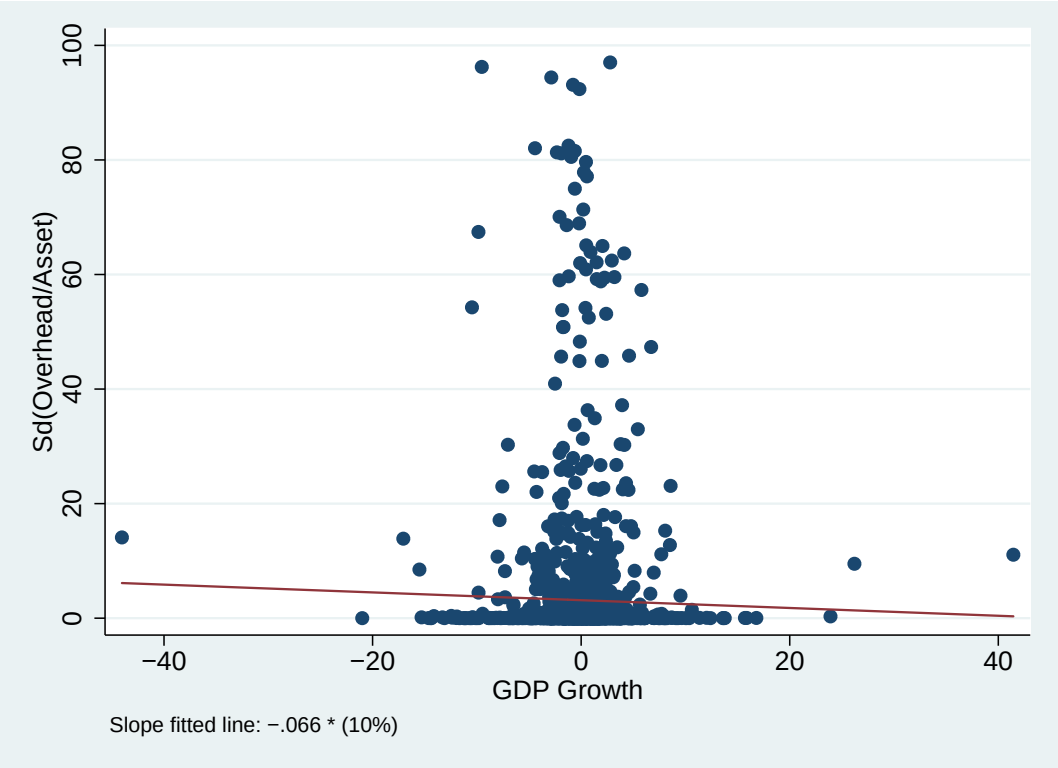


Figure 4: Heterogeneous banking efficiency maps into lower GDP growth



3.1 The heterogeneous entrepreneurship sector

A firm, set up by an entrepreneur, can only have recourse to a single financial intermediary to get a financing ; the type of firms I focus on here is small and medium sized firms that cannot issue debt directly on the stock markets.

Entrepreneurs can undertake three types of project, High, Medium or Low, which are associated with a high, medium or low probability of success, respectively p_H, p_M and p_L, with $p_H > p_M > p_L > 0$. Denote $\Delta_H = p_H - p_M$ and $\Delta_M = p_M - p_L$. For a given productivity level α and interest rate r, both H and M projects have a positive NPV, while project L has a negative NPV.

Assumption 1. $\alpha p_H > \alpha p_M > r > \alpha p_L$

Because of the presence of the negative NPV project, I have the classical source of moral hazard which forces banks to provide an adequate loans contract in order to prevent entrepreneurs from diverting part of the resources supposed to be invested in the firm in order to reap a private benefit b.

Then the simultaneous existence of two positive NPV projects creates an other source of moral hazard, which is the natural counterpart of having an heterogeneous entrepreneurship

sector. I consider a continuum of entrepreneurs indexed by their effort cost e per unit of capital invested in the firm; for simplicity the density function $f(e)$ is assumed to follow a uniform distribution $e \to \mathrm{U}([e^{min}; e^{max}])$. Thereafter in the paper, I use e as the cost of effort of some entrepreneur picked in the previous distribution. Thus entrepreneurs face the following moral hazard problem: if they choose to undertake a high yield project H, they will have to bear a pecuniary cost of effort e if induced by bank monitoring, or $e^l = e - l < e$ if the entrepreneur decides to provide efforts without being scrutinized by the monitor. However, if they choose to undertake a medium yield project M, the cost of effort will be γe, i.e. proportionally less than the entrepreneurs' type if the maximum effort was provided; for simplicity, I set $\gamma = 0$ as it allows to derive closed formed solutions. Effort cost depending on the project undertaken is given by:

$$E_j = \begin{cases} e, & \text{for j=H, then proba of success } p_H \\ 0, & \text{for j=M, then proba of success } p_M \\ -b, & \text{for j=L, then proba of success } p_L \end{cases}$$

Then entrepreneurs will default with a positive probability $1 - p_j$, $j \in \{H, M, L\}$, and I assume that everything is lost in case of default. Henceforth the capital needed by an entrepreneur to run his firm has to be borrowed at a higher cost than the market risk-free interest rate. Entrepreneurs will first invest their own wealth A, which is homogeneous across all of them, and borrow B, which will depend on entrepreneurs' type, to run a project of size $k = A + B$. Note that, by definition, for the same positive NPV project and a linear production function, an entrepreneur will seek to contract with the bank offering the largest loan.

The profit of an entrepreneur is expressed as follows:

$$\pi_j = p_j(y^i - R^i B^i) - E_j k^i \tag{1}$$

with $k^i = B^i + A$, $i \in \{h, l\}$ the type of bank and $j \in \{H, M, L\}$ the type of project undertaken, with high or low effort. $y = \alpha k^i$ is a constant return to scale production function with productivity parameter α; the price of the output is normalized to 1.

Due to the presence of fixed costs $E_j A$ and constant returns $p_j \alpha A$, if, on average, the entrepreneur can be better off making effort and behaving, the marginal gain of making effort (behaving) is lower than shirking (diverting funds). Nevertheless, the project is not separable: if the entrepreneur decides to shirk, it shirks for the whole project, likewise if he behaves or not.

Assumption 2. $e > \Delta_H(\alpha - R^i)$

Assumption 3. $b > \Delta_M(\alpha - R^i)$

Note that only the H-project can sustain effort $\alpha p_H - e > r$, while one would never choose to make effort if the probability of success was to remain medium, that is to say $\alpha p_M - e < r$. Nevertheless, from a social point of view, project H would be preferred as its NPV adjusted for the pecuniary cost of effort is higher $\alpha p_H - e - c > \alpha p_M$, that is to say the marginal product of effort is larger than the marginal product of shirking. Likewise, it is always socially optimal to avoid the negative NPV project, since the marginal product of behaving is assumed larger than the marginal product of diverting funds $p_M(\alpha - r) > p_L(\alpha - r) + b$.

Assumption 4. $\Delta_M(\alpha - r) > b$

3.2 The heterogeneous banking sector

I introduce some heterogeneity in the banking sector, in the form of a different cost structure. I consider two competing banks with different monitoring cost, c^l for the bank featuring a higher monitoring ability and low cost of doing so, and c^h for the bank featuring a lower monitoring ability and a high associated cost, that is to say $c^h = (1 + \phi)c^l$ with $\phi > 0$ being the degree of bank heterogeneity in monitoring abilities. The Bertrand duopoly approach is only necessary in order to depart from the fully competitive case and get positive profits for banks in equilibrium.

Each competing financial institution will design two loan contracts, respectively $C^h(R^h, B^h)$ and $C^l(R^l, B^l)$, for a given level of personal contribution A. The type of the entrepreneur is not observed at the contract design stage, but the bank knows that some entrepreneurs will have to be compensated for their effort and factors that in to design a contract from which the entrepreneur will not deviate. Henceforth, each bank will design two types of contracts for entrepreneurs making effort (H project) and those making less effort or shirking but who do not misbehave (M project). In addition, when designing its set of contracts, each bank will take into account the fact that entrepreneurs will compare the different loan contracts, so that banks engage in ex-ante contract competition with each other before finalising their contracts. This seems realistic as banks usually fix their credit standard every month by taking into account both their financing ability as well as the situation of competitors, and on a day-to-day basis they only adjust generic contracts to each borrower, depending on the assessment of the type of the borrower.

To finance their loans, banks have to levy funds on the capital markets at the risk-free interest rate r; their role is thus to pool idiosyncratic risks associated with individual loans in order to offer a constant return on savings that does not depend on the probability of success of the firms. Thus a bank will offer a loan contract that covers both the cost of financing r (or opportunity cost of providing a loan), and the monitoring cost c if it decides to monitor.

$$\Pi^i = p_j R^i B^i - c^i k^i - r B^i \tag{2}$$

with $i \in \{h, l\}$. So it has to be that the bank finds it profitable to make at least H-type contracts, that is to say for a marginal unit of loan I have $p_H R^i > c^i + r$. Moreover, I consider only the meaningful case where the monitoring cost is lower than the cost of financing a loan $c^i < r$.

Assumption 5. $p_H R^i > c^i + r$ *with* $r > c^i$

Note that I abstract from differences on the liability side between banks having different cost structures on the asset side, in order to isolate the effect of bank monitoring heterogeneity both on credit allocation and credit fluctuations, and show that despite the fact that one type of bank is clearly dominated –higher monitoring costs are not compensated for instance by easier access to capital markets–, it may have non trivial implications in equilibrium in a model with entrepreneurs heterogeneity.

3.3 Information

Ex-ante, the banks do not know the specific characteristics of the entrepreneurs –neither e nor b– but know each distribution among the pool of entrepreneurs. Ex-post, after the contracts have been designed by the banks, for each entrepreneur who enters a bank, it immediately reveals its private benefit from shirking and diverting funds b, and, upon monitoring by the bank, its effort cost is revealed.

Hence the social benefit in having financial intermediaries lies in their ability to perfectly screen out bad projects with negative NPV. If one was to allow financing by outside experts, they would not be able to monitor and discriminate depending on the cost of effort, hence leading to a pooled equilibrium where outside financiers would treat everyone like the worst agent. This case is left as an extension.

Note that contract commitment is not strictly speaking necessary in equilibrium as the incentives are such that banks would not find it profitable to deviate.

The time-line of the model is displayed in figure 5.

Definition 1. *An equilibrium is defined by a set of two active contracts*

$$\zeta = \left\{ C^{*i}_{EC} | e \leq \bar{e}, C^{*i'}_{non-EC} | e > \bar{e} \right\}$$

with $i, i' \in \{h, l\}$*, one on each segment of the market, monitored (effort compatible for high yield projects) or not (non-effort inducing for medium yield projects), denoted respectively* $\{EC, non-EC\}$*; market segmentation is given by the endogenous effort cost threshold* $\bar{e}$*. The set of active contracts is chosen by entrepreneurs among a set of four contracts* ξ *such that, for high (medium) yield project* H *(*M*):*

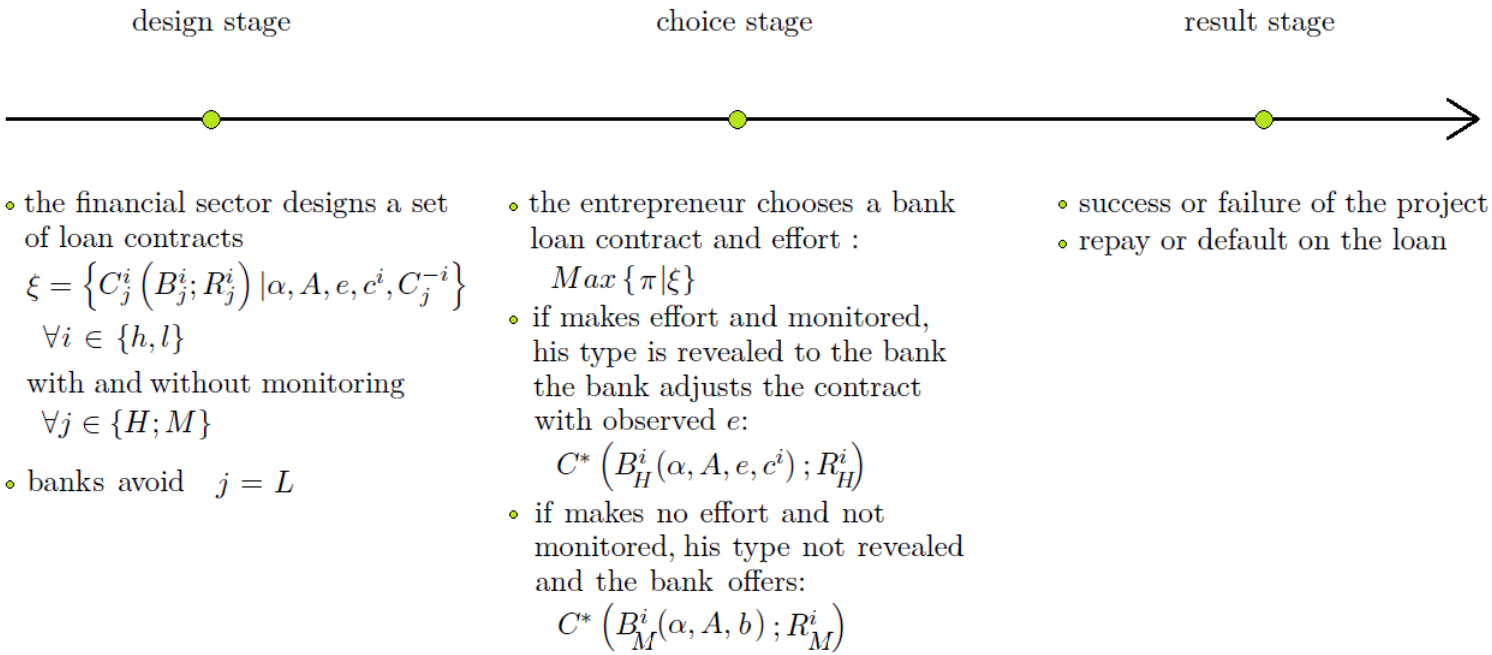
$$\begin{aligned} \left\{ C^{*i}_{EC} | e \leq \bar{e} \right\} &= argmax_{\{B \geq 0, R > 0\}} \{\pi_H | \xi\} \\ \left\{ C^{*i'}_{non-EC} | e > \bar{e} \right\} &= argmax_{\{B \geq 0, R > 0\}} \{\pi_M | \xi\} \end{aligned}$$

with

$$\xi = \{C^h_{EC}, C^l_{EC}, C^h_{non-EC}, C^l_{non-EC}\}$$

where each contract proposed by an institution $i \in \{h, l\}$ *on a segment* $j \in \{EC, non - EC\}$ *is defined by the pair* $C^i_j(B^i_j, R^i_j | \alpha, r, A, e, c^i, C^{-i}_j)$*.* B^i_j *stands for the loan size,* R^i_j *for the interest rate offered,* α *for the productivity,* r *for the market interest rate,* A *for the level of personal contribution,* e *for the type of the entrepreneur,* c^i *for the monitoring cost and* C^{-i}_j *for the banking contract designed by the competitor.*

Figure 5: Time-line of the decision process



4 Equilibrium without bank heterogeneity

The fully competitive equilibrium, which would arise for instance by assuming a continuum of banks of each type h and l, is in fact equivalent to the Bertrand duopoly case without bank heterogeneity ($\phi = 0$), that is to say 2 or more completely identical banks [6] competing to design the most attractive loan contracts. In this benchmark case, all rents are transferred to the entrepreneurs and banks make zero profits, that is to say there is no positive markup in the banking industry.

Since there exists two types of projects which can be undertaken, High (H) or Medium (M) yield, the setting of the model *de facto* implies that there will be a market segmentation with some entrepreneurs choosing one project or the other. Thus banks need to design one contract supporting each project by making sure the entrepreneur will not deviate from it.

4.1 Contract design stage : High yield compatible

The contract supporting the High yield project has to be such that the Effort Compatibility (EC) constraint (3), which ensures the largest positive NPV project is undertaken by the entrepreneur, and the Incentive Compatibility (IC) constraint (4), which ensures the entrepreneur does not pick the negative NPV project, are satisfied. So each bank has to make sure that the following holds:

$$p_H(\alpha k - RB) - ek \geq p_M(\alpha k - RB) \tag{3}$$

$$p_H(\alpha k - RB) - ek \geq p_L(\alpha k - RB) + bk \tag{4}$$

But then the bank has an incentive to provide an H-type contract only if it is more profitable to learn the type of the entrepreneur and monitor his effort. As a consequence the bank incurs a cost c, such that its Monitoring Constraint (MC) and Participation Constraint (PC) are respectively given by:

$$p_H RB - ck \geq p_M RB \tag{5}$$

$$p_H RB - ck \geq rB \tag{6}$$

Equation (6) states that the expected return from offering an H-type contract has to be greater

6. Thus the subscript $\{h, l\}$ is dropped for the homogeneous case.

or equal to the opportunity cost of lending on the market.

The fully competitive H-type contract is obtained as follows; the equilibrium loan size B^* is obtained from equations (3) and (4) which can be rearranged as follows:

$$B \leq \frac{\alpha - e/\Delta_H}{e/\Delta_H + R - \alpha} A \equiv B_{EC} \tag{7}$$

$$B \leq \frac{\alpha - \frac{b-e}{p_H - p_L}}{\frac{b-e}{p_H - p_L} + R - \alpha} A \equiv B_{IC} \tag{8}$$

Equations (7) and (8) give the maximum loan amount such that entrepreneurs have respectively no incentives to make less effort and no incentives to divert part of the resources. For the problem of the entrepreneur to make sense, it has to be that we have $B_{EC} < B_{IC}$, that is to say that when the entrepreneur is better off making the maximal efforts, he does not find it profitable divert funds, while not diverting funds may not necessarily imply the maximal effort-making. So I focus on the case where the following restriction holds.

Assumption 6. $e > b\frac{\Delta_H}{2p_H - p_M - pL}$

Since the entrepreneur's profit is increasing in k, he will seek for the contract that allows for the maximum leverage, so in order to remain competitive, each bank will have to offer the upper bound on borrowing:

$$B_H^c = B_{EC} \tag{9}$$

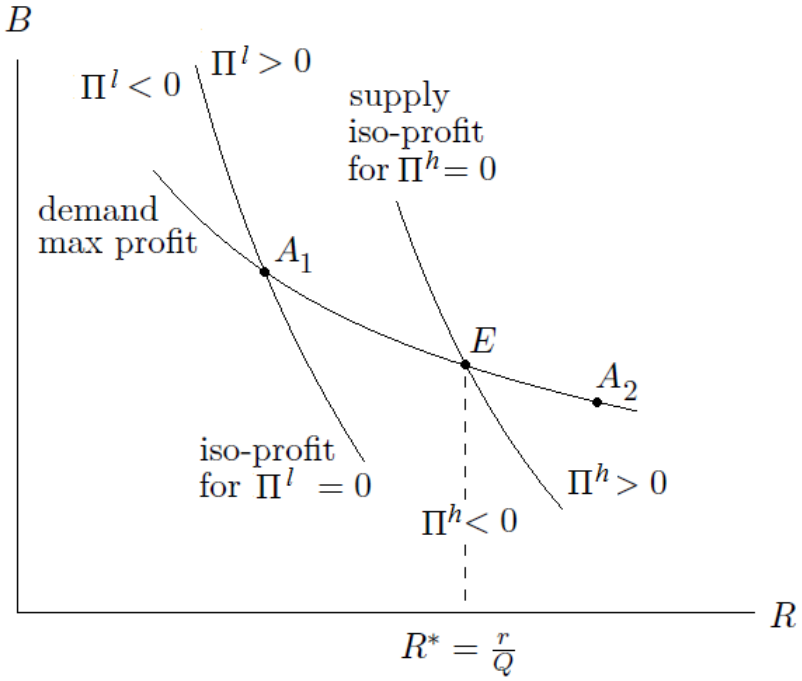
Then the equilibrium interest rate R_H^c is obtained from equation (6) which must hold as an equality due to perfect competition and the associated zero profits condition obtained from equating equation (2) to zero. Rearranging and plugging in the value of B_H^c given in equation (9), I obtain the following competitive equilibrium interest rate:

$$R_H^c = \frac{r}{Q^c} \tag{10}$$

where $Q^c = p_H - \frac{c\Delta_H}{\alpha\Delta_H - e}$ can be thought of as the risk premium adjusted for competition. Since $Q < p_H$, banks offer an interest on loans which comprises a risk premium over the risk free rate and is also adjusted for the cost of monitoring.

Due to this ex-ante contract competition, the model features countercyclical markup $\frac{1}{Q}$ when a productivity shock hits. R_H^c is positive under the condition that $c < p_H(\alpha - e/\Delta)$ which states

Figure 6: H-type contract indifference curves



that the expected marginal return of an entrepreneur making effort has to be larger than the marginal cost of monitoring to ensure he does not shirk.

Assumption 7. $c < p_H(\alpha - e/\Delta_H)$

Proposition 1. *Without bank heterogeneity or equivalently full competition, all banks offer the same H-type-inducing contract* $C_H^c\left(R_H^c, B_H^c\right)$ *which successfully mitigates the two sources of moral hazard, with:*

$$\begin{aligned} B_H^c &= B_{EC} \\ R_H^c &= r/Q^c \end{aligned}$$

under assumptions 1-2 and 5-7.

Figure 6 represents the contract on the sub-segment of the market successfully implementing the High yield projects. The demand curve of loans by entrepreneurs corresponds to equation (7), that is to say the maximum leverage allowed for entrepreneurs making the effort. The equilibrium pair $\{B; R\}$ lies on this demand curve since entrepreneurs are better off the larger the loan size. The intuition for the shape of the iso-profit curves of the bank is as follows: the bank is indifferent between making a small loan but earning a large interest margin and making a large loan but only making a small margin on each unit lent. In the absence of bank

heterogeneity and thus perfect competition, all banks would be as efficient as the others (or be like the most efficient one, bank l for low cost) and the zero profit equilibrium for the bank would have been reached in point $A1$.

4.2 Contract design stage : Medium yield compatible

The contract supporting the Medium type of project is such that the bank is better off not knowing the type of the entrepreneur; that is to say it accepts a lower expected probability of success as it saves on the cost of monitoring necessary to know the entrepreneur's type. But then the bank cannot condition loan size on entrepreneurs type, such that they will be better off leveraging the maximum they can, which will fail to give them the appropriate incentives to make effort.

At the design stage, the bank will take into account the fact that the entrepreneur, at the choice stage, will participate without providing efforts while the entrepreneur is still better off not diverting funds away; both conditions, that the EC does not hold while the IC does, are respectively given by:

$$p_H(\alpha k - RB) - e^l k \leq p_M(\alpha k - RB) \tag{11}$$

$$p_M(\alpha k - RB) \geq p_L(\alpha k - RB) + bk \tag{12}$$

From (11), one gets a lower bound $B_{EC} < B^*$ above which the entrepreneur is better off not making too much efforts. Then a bank $i \in \{h, l\}$ still participates but is better off not learning the entrepreneur cost of effort which saves on its own cost; so the MC constraint does not hold while the PC does, both constraints being respectively given by:

$$p_H RB - ck \leq p_M RB \tag{13}$$

$$p_M RB \geq rB \tag{14}$$

Since the banks face perfect competition and do not differ from each other on this segment of the market (no monitoring cost is paid), the equilibrium will be symmetric and lie at the zero profits constraint of the banks. Thus the PC of the bank in (14) will be binding, so that the interest rate offered for the M-yield supporting contract R_M^c will just be the risk-free rate adjusted for the default premium.

Using the above-mentioned profitability assumptions, the set of constraint (11), (12) and (13) for an institution $i \in \{h, l\}$ can be expressed as:

$$B \geq \frac{\alpha - e^l/\Delta_H}{e^l/\Delta_H + R - \alpha} A \equiv \bar{B}_{EC} \tag{15}$$

$$B \leq \frac{\alpha - b/\Delta_M}{b/\Delta_M + R - \alpha} A \equiv \bar{B}_{IC} \tag{16}$$

$$B \leq \frac{c}{\Delta_H R - c} A \equiv \bar{B}_{MC} \tag{17}$$

Since entrepreneurs are better off leveraging the most they can, the loan size will be given by one of the two potential upper bound of equation (16) and (17). Note that equation (17) is necessarily met provided that (15) holds. Indeed, assume that (13) is not met, so that I have $p_H RB - ck \geq p_M RB$. But equation (15) implies that the entrepreneur is better off shirking so that such level of loans is not effort compatible: even if the bank was to pay the cost to learn the type e of the entrepreneur, at such loan level, it could not use this information to implement an effort compatible contract. So the realized probability of success would be lower at p_M; the incentive constraint of the bank would thus write $p_M RB - ck \geq p_M RB$ which is obviously violated. So the bottom line is that equation (17) cannot be deviated from if (15) holds; since equation (15) is a lower bound it is necessarily met in equilibrium, so that the loan size offered by the bank will be given by the incentive compatible equation (16) holding as an equality.

Again on this segment of the market, I restrict to the meaningful case where $\bar{B}_{IC} \geq \bar{B}_{EC}$.

Assumption 8. $b \geq \frac{\Delta_M}{\Delta_H} e^l$

Proposition 2. *Without bank heterogeneity or equivalently full competition, all banks offer the same M-type-inducing contract* $C_M^c\,(R_M^c, B_M^c)$ *which successfully mitigates the traditional source of moral hazard by avoiding the negative NPV project but fails to mitigate the second source of moral hazard by not supporting the highest yield project, with:*

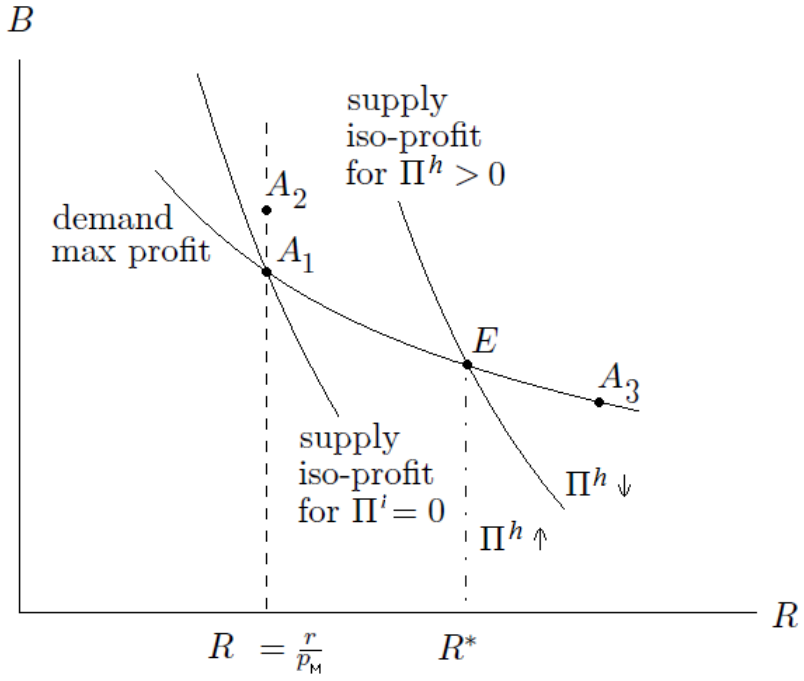
$$B_M^c = \bar{B}_{IC}$$

$$R_M^c = r/p_M$$

under assumptions 1, 3, 4 and 8.

Figure 7 represents the equilibrium contract on the M-yield segment of the market. The

Figure 7: M-type contract indifference curves



dotted line represents the interest rate such that any bank i breaks even. The feasibility set for firms is represented by the area under the curve which maximises entrepreneurs' profits which represents equation (16). So the competitive equilibrium is given by point A_1, while point A_2 is not achievable as the banks would never finance negative NPV projects.

4.3 Contract choice stage : separating equilibrium

An entrepreneur e will self-select the M-yield project if it is more profitable for him not to make efforts, that is to say if he is better-off choosing the contract $C_M^c(R_M^c, B_M^c)$, so that :

$$\begin{aligned} C_M^c(R_M^c, B_M^c) &\succ C_H^c(R_H^c, B_H^c) \\ \pi_M(R_M^c, B_M^c) &\geq \pi_H(R_H^c, B_H^c) \\ p_M\left(\alpha(B_M^c + A) - R_M^c B_M^c\right) &\geq p_H\left(\alpha(B_H^c + A) - R_H^c B_H^c\right) - ek_H^c \end{aligned} \tag{18}$$

The cutoff level of effort $\bar{e}^c$ is obtained when the last equation holds with equality; hence the proposition follows.

Proposition 3. *The separating competitive equilibrium in the absence of bank heterogeneity ($\phi = 0$) is such that:*

- *entrepreneurs with a low cost of effort* $e < \bar{e}^c$ *will undertake the H-yield project and choose contract* $C_H^c\left(R_H^c, B_H^c\right)$*;*
- *entrepreneurs with a high cost of effort* $e > \bar{e}^c$ *will undertake the M-yield project and choose contract* $C_M^c\left(R_M^c, B_M^c\right)$*.*

5 Social optimum

I now turn to the description of the optimal contracts which would maximise aggregate surplus. The optimal outcome is to maximise leverage on the H-yield contracts since I have a larger marginal return of leverage $\alpha p_H - e - c > \alpha p_M$; so entrepreneurs with cost of effort below $\alpha\Delta_H - c \equiv \bar{e}^o$ should be incentivised to choose the H-type projects. So the optimal benchmark is given by equilibrium contract A_1 on graph 6 and around A_3 on graph 7 in order to make sure the entrepreneurs who can make effort are not attracted on the M-yield segment of the market. Thus the optimum fully takes into account the disincentive effect induced by the possibility of switching between different market segments inherent to a model with more than one positive NPV project; the externality induced by adverse selection by entrepreneurs across markets is internalized.

Proposition 4. *The social optimum which fully internalises the effect of adverse selection in a segmented market is given by the pair of contracts:*

- *entrepreneurs with a low cost of effort* $e < \bar{e}^o$ *will undertake the H-yield project and choose contract* $C_H^o\left(R_H^o = R_H^c, B_H^o = B_H^c\right)$*;*
- *entrepreneurs with a high cost of effort* $e > \bar{e}^o$ *will undertake the M-yield project and choose contract* $C_M^o\left(R_M^o > R_M^c, B_M^o = B_M^c\right)$*;*

where $\bar{e}^o > \bar{e}^c$*.*

6 Equilibrium with two heterogeneous banks

I now investigate the market equilibrium case with bank heterogeneity in monitoring abilities, that is to say one l bank with low cost of monitoring and one h bank with high monitoring costs where $c^h = (1+\phi)c^l$ with $\phi > 0$.

6.1 Contract design stage : High yield compatible

The contract supporting the High type of project has to be such that the Effort Compatibility (EC) constraint (3) and the Incentive Compatibility (IC) constraint (4) of the entrepreneur as well as the Monitoring Constraint (MC) (5) and Participation Constraint (PC) (6) of the bank is satisfied. Using the profitability assumptions, the set of four constraints for an institution $i \in \{h, l\}$ can be expressed as:

$$B^i \leq \frac{\alpha - e/\Delta_H}{e/\Delta_H + R^i - \alpha} A \equiv B_{EC} \tag{19}$$

$$B^i \leq \frac{\alpha - \frac{b-e}{p_H - p_L}}{\frac{b-e}{p_H - p_L} + R^i - \alpha} A \equiv B_{IC} \tag{20}$$

$$B^i \geq \frac{c^i}{\Delta_H R^i - c^i} A \equiv B^i_{MC} \tag{21}$$

$$B^i \geq \frac{c^i}{p_H R^i - c^i - r} A \equiv B^i_{PC} \tag{22}$$

Equations (19) and (20) give the maximum loan amount such that entrepreneurs have respectively no incentives to provide the minimum level of effort and no incentives to divert part of the resources. Recall that from assumption 6, I have $B_{EC} < B_{IC}$ for the problem to be meaningful. Equations (21) and (22) give a minimum loan amount as the bank, in order to be incentivised and make profits, has to cover at least the fixed cost of monitoring, proportional to the size of collateral/entrepreneur personal investment $c^i A$.

The equilibrium H-type contract is obtained as follows: at the design stage, the bank takes into account the future choice of the entrepreneur; if the entrepreneur makes the required effort, he can still compare the H-type contract offered by the high and low ability bank. So each bank anticipates on the competitive implication of the cost of its own contract.

Assume for the moment that I indeed have :

$$\text{E1} : \quad B_{EC} \geq B^i_{MC}$$

$$\text{E2} : \quad B_{EC} \geq B^i_{PC}$$

Since the entrepreneur's profit is increasing in k, he will seek for the contract that allows for the maximum leverage, so in order to remain competitive, the bank will have to offer the upper

bound on borrowing:

$$B^* = B_{EC} \tag{23}$$

If both the high and low ability bank allow for the maximum leverage demanded by the entrepreneur, entrepreneurs will choose the contract offering the lowest interest rate. So ex-ante, the high-ability bank will design a contract that offers an interest rate which would be low enough to drive the low-ability bank out of the market; recall that the latter is defined by a less efficient cost structure when operating on the segment of the market that requires monitoring, so that the more efficient bank will be able to extract a surplus by offering a limit interest rate, higher than the fair pricing of default risk. From the profitability condition equation (6) for the two banks, I have the following reaction functions:

$$\frac{\Pi^l + c^l(B^* + A) + r(R_n^h)B^*}{p_H B^*} = R_{n+1}^l \tag{24}$$

$$\frac{\Pi^h + c^h(B^* + A) + r(R_n^l)B^*}{p_H B^*} = R_{n+1}^h \tag{25}$$

In addition, I know that the market interest rate $r(R^*)$ decreases with the interest on the loan, since larger interest repayment on loans decreases the amount borrowed ($\frac{\partial B^*}{\partial R^i} < 0$), and lower refinancing need by banks reduce the pressure on the capital market and decreases their financing cost r. So for bank i to remain on the market after bank k lowered the interest on its loan, interest repayment must be realigned such that $R_{n+1}^i < R_n^{k\backslash i}$ for $i, k \in \{l, h\}$. So in equilibrium, the bank most efficient in monitoring will reach $\Pi^h = 0$ while $\Pi^l > 0$. Rearranging equation (25) and plugging in the value of B^* given in equation (23), I obtain the following competitive equilibrium interest rate:

$$R^* = \frac{r}{Q} \tag{26}$$

where $Q = p_H - \frac{c^l(1+\phi)\Delta_H}{\alpha\Delta_H - e}$ can be thought of as the risk premium adjusted for competition so that the most efficient bank generates a positive net interest margin which features a countercyclical markup. R^* is positive under assumption 7 for $c = c^{pu}$.

As a consequence, one can verify that equation (E1) holds with an inequality while equation (E2) holds with equality for the least efficient bank and with an inequality for the most efficient bank.

Proposition 5. *With two heterogeneous banks, at the effort compatible equilibrium, H-type projects are financed by the bank with low monitoring cost c^l which offers a contract $C_H^l\left(R_H^*, B_H^*\right)$ which successfully mitigates the two sources of moral hazard, with:*

$$\begin{aligned} B_H^* &= B_{EC}\left(\phi\right) \\ R_H^* &= r/Q\left(\phi\right) \end{aligned}$$

under assumptions 1, 2, 5, 6 and 7.

The competitive equilibrium with heterogeneous banking is depicted in figure 6 by point E, that is to say when the least efficient bank (bank h) with higher monitoring costs c^h is driven out of the market and makes no profit as a result of the competition in contracts.

6.2 Contract design stage : Medium yield compatible

Now that banks have heterogeneous monitoring abilities, a rent can be extracted by the most efficient one on the H-segment of the market. In principle the M-segment of the market is such that both banks make zero profits as they are perfectly competitive since banks have the same opportunity cost of lending (the heterogeneity in monitoring is not relevant for M-projects). But instead of competing in the M-segment of the market, the most efficient bank which makes profits on the H-segement of the market will seek to engage on the M-segment of the market only if does not reduce its profitability. So the design of the M-contract will be subject to a Profits-Preserving (PP) constraint.

Indeed, with a fairly priced M-contract, it is relatively more attractive than the H-contract for some entrepreneurs; but the efficient bank, by allowing some degree of market power to the less efficient bank on the M-segment of the market, would ensure that the M-contract, without competitive pricing, is now relatively less attractive than the H-contract for some entrepreneurs. Henceforth, more entrepreneurs would choose H-yield projects, which means more profits for the efficient bank if it relaxes the competition on the M-segment of the market.

Then the bank with higher monitoring costs will benefit from the market power it gets on the M-segment of the market and thus will to maximise its now positive profits which are given

by the following equation:

$$\begin{aligned}\Pi^{h*}_{total} &= \int_{\bar{e}^*}^{e^{max}} \Pi^h \mathtt{d}e \\ &= \int_{\bar{e}^*}^{e^{max}} \left(p_M R^*_M B^*_M - r B^*_M\right) \mathtt{d}e \\ &= \left(e^{max} - \bar{e}^*\right)\left(p_M R^*_M - r\right) B^*_M \end{aligned} \tag{27}$$

To pin down the equilibrium interest rate associated with the rent extraction, the less efficient bank faces the following trade-off which : by increasing its market power, it extracts more rent per unit of loan (second term in equation (27)), but the size of each individual loan decreases (third term) and the number of customers choosing M-projects decreases (first term) as H-supporting contracts become relatively more attractive.

Proposition 6. *With two heterogeneous banks, at the non-effort –but incentive– compatible equilibrium, M-type projects are financed by the low efficiency bank which offers a contract* $C^h_M(R^*_M, B^*_M)$ *that successfully prevents negative NPV projects but fails to provide incentives to make efforts, with*

$$\begin{aligned} B^*_M &= \bar{B}_{IC} \\ R^*_M &= \underset{R>r/p_M}{\texttt{argmax}}\, \Pi^h_{total} \end{aligned}$$

under assumptions 1, 3, 4 and 8, and the two sufficient conditions that e^{max} *be small enough and the NPV of the M-project not too high.*

The competitive equilibrium on the M-yield segment of the market with heterogeneous banking is depicted in figure 7 by point E. If the less efficient bank (bank h for high costs) lends at a point at the right of A_1, then it moves from the competitive case to the case with a positive markup; by moving from point A_1 towards E, the less efficient bank decreases the attractiveness of its M-contract and *de facto* increases the profits of the efficient bank on the H segment of the market. Beyond point E, the less efficient bank increases so much its market power that it starts making less profits as entrepreneurs do not undertake M-projects any more or reduce the size of their projects.

6.3 Contract choice stage : separating equilibrium

As in the case without bank heterogeneity, once contracts have been competitively designed by the banking sector, entrepreneurs will self-select bank loan contracts in order to maximise their profits.

Proposition 7. *The seperating competitive equilibrium in the presence of bank heterogeneity ($\phi > 0$) is such that:*

- *entrepreneurs with a low cost of effort $e < \bar{e}^*$ will undertake the H-yield project and choose the contract provided by the bank with low screening costs profile $C_H^{l*}(R_H^*, B_H^*)$;*
- *entrepreneurs with a high cost of effort $e > \bar{e}^*$ will undertake the M-yield project and choose the contract provided by the bank with high screening cost profile $C_M^{h*}(R_M^*, B_M^*)$.*

7 Equilibrium with several heterogeneous banks

I now move beyond the specific duopoly setting by introducing several banks heterogeneous in their loan monitoring efficiency. I still denote by l the bank with the lowest cost of monitoring but now the inefficiency wedge ϕ^b for each bank $b \in \{1, ..., n\}$ is such that $\phi^b \in \{0, \phi^2, ..., \phi^n\}$ and $c^b \in \left\{c^1 = c^l, c^2, ..., c^n\right\}$.

7.1 Contract design stage : High yield compatible

The same story goes through, that is to say the bank l with the largest comparative advantage in monitoring loans will take over the whole H-yield market segment and will design a contract that drives out of the market all its competitors while retaining the largest market power it can. Thus the equilibrium contract will be pinned down by the competition with bank $b = 2$ which is the most efficient in screening loans among the banks $b \in \{2, ..., n\}$ that have an inefficiency wedge.

Thus proposition 6 carries out to the n bank case where the equilibrium contract is given by the distance $\phi^{b=2}$ in terms of efficiency between the first and the second bank. So the limit case where $n \longrightarrow \infty$ is down to proposition 1 with perfect competition.

Proposition 8. *With several heterogeneous banks, at the effort compatible equilibrium, H-type projects are financed by the bank with the lowest monitoring cost c^l which offers a contract*

$C_H^l\left(R_H^{**}, B_H^{**}\right)$ *which successfully mitigates the two sources of moral hazard, with:*

$$\begin{aligned} B_H^{**} &= B_{EC}\left(\phi^{b=2}\right) \\ R_H^{**} &= r/Q\left(\phi^{b=2}\right) \end{aligned}$$

under assumptions 1, 2, 5, 6 and 7.

7.2 Contract design stage : Medium yield compatible

On the M-yield segment of the market, the situation differs since now $n > 2$ banks have the same opportunity cost of lending when no monitoring is undertaken, but only the most efficient bank $b = 1$ has a profits preserving constraint. Nevertheless, by refraining from entering the M-yield segment of the market, this latter bank cannot really impact the competitive pressure on this segment as it is only one bank among n competitors. Thus two cases have to be distinguished.

7.2.1 No markup within banking segments

So far the model allowed to derive an endogenous markup across market segments, and in the specific duopoly case in each segment of the market. With n heterogeneous banks, the markup across sector remains and vanishes as the number of banks tends to infinity. But the markup in the specific M-yield segment of the market disappears if the n banks competes fully, so that the equilibrium contract is given, for the interest rate by the fair risk-adjusted interest rate, and for the loan size by the incentives to undertake the negative NPV project.

7.2.2 Presence of a markup within banking segments

Alternatively, empirical evidences show that the banking sector is not fully competitive even within specific lines of business, especially when a small number of banks operate in these segments. Thus instead of deriving an endogenous market power intensity as in the duopoly case, the markup ψ would be exogenously given, with $\psi = 0$ being a specific case without markup in the M-yield segment of the market. The case $\psi > 0$ can be thought of as a result of Hotelling spacial costs or monopolistic competition with complementarities in banking services. The following proposition encompasses the two cases.

Proposition 9. *With several heterogeneous banks, at the non-effort –but incentive– compatible equilibrium, M-type projects can be financed by any bank b where each of them offers the same contract* $C_M^b(R_M^{**}, B_M^{**})$ *that successfully prevents negative NPV projects but fails to provide incentives to make efforts, with*

$$\begin{aligned} B_M^{**} &= \bar{B}_{IC} \\ R_M^{**} &= \frac{r}{p_M} + \psi \end{aligned}$$

under assumptions 1, 3, 4 and 8, and the two sufficient conditions that e^{max} *be small enough and the NPV of the M-project not too high.* ψ *is the intensity of the markup in this market segment depending on the market structure considered.*

In figure 7, the equilibrium on the M-yield segment of the market with several heterogeneous banks lies between points A_1 and A_3 where a larger market power moves the equilibrium away from A_1.

7.3 Contract choice stage : market segmentation

With several banks heterogeneous in their monitoring efficiency, the market segmentation remains, driven by the set of productive choices H, M, L available to entrepreneurs. However, the equilibrium is only partly separating, with the most efficient bank absorbing the whole market for high yield entrepreneurs, but all banks are pooled on the M-segment of the market and an entrepreneur can match randomly with any bank.

Proposition 10. *The endogenously segmented competitive equilibrium in the presence of several heterogeneous banks is such that:*

- *entrepreneurs with a low cost of effort* $e < \bar{e}^{**}$ *will undertake the H-yield project and choose the contract provided by the most efficient bank* $C_H^{l**}(R_H^{**}, B_H^{**})$*;*
- *entrepreneurs with a high cost of effort* $e > \bar{e}^{**}$ *will undertake the M-yield project and match randomly with any bank offering the contract* $C_M^{b**}(R_M^{**}, B_M^{**})$*.*

8 Distortions associated with banking heterogeneity

Although the model presented here is by nature static, the impact of banking heterogeneity can be analysed along two dimensions, which would correspond, in the terminology of macroe-

conomists, to long run distortions and short run fluctuations, that is to say comparative statics of the levels of aggregate variables and of the deviations after a productivity shock. Note that the modeling strategy employed here does not allow for a quantitative analysis as it assumes a specific distribution of entrepreneurs which determines endogenous market shares across bank efficiency.

8.1 Allocative distortions associated with banking heterogeneity shocks

When the intensity of banking heterogeneity increases, that is to say the difference of bank efficiency in monitoring abilities widens, it increases the rent the most efficient bank can extract on high ability entrepreneurs, so that it is relatively less attractive for them to remain on the H-yield, effort-compatible segment of the market, which de facto increases the pool of shirking entrepreneurs who would have otherwise provided effort. Thus less effort is provided, so that less high-yield projects are undertaken, which reduces realised aggregate productivity further compared to the potential aggregate productivity at the social optimum.

Proposition 11. *The larger the banking sector heterogeneity in bank monitoring efficiency between the most efficient and other banks, the stronger the adverse selection towards less productive projects allowing for shirking.*

Corollary 1. *A shock on the distribution of bank monitoring efficiency –i.e. an increase in banking heterogeneity between the most efficient and other banks– reduces average entrepreneurial productivity.*

When looking at the competitive homogeneous equilibrium versus the heterogeneous banking equilibrium, two effects on incentives and effort-making work in opposite directions. By introducing a less efficient bank, you decrease competition on the H-segment of the market and create a rent for the efficient bank on this segment, which should lower incentives to undertake efforts. In the meantime, the introduction of the less efficient bank generates a profits-preserving behaviour for the efficient bank which increases the market power of the less efficient bank on the M-segment of the market; this should result in lower incentives to shirk and de facto more incentives to turn to effort-making contracts.

If the intensity ψ of the market power on the M-segment by the less efficient bank is too small, that is to say if the financing condition of the M-contract become too attractive, then there is an excessive number of entrepreneurs choosing not to undertake efforts. In the meantime, if

the heterogeneity in bank efficiency ϕ is too high, then having less efficient banks allows for an excessively high rent to be extracted on the H-segment of the market which will discourage effort. So the worst case in terms of the distortions induced by banking heterogeneity on effort making, which channels via the adverse selection process, would be a very high level of heterogeneity in monitoring efficiency and high market competition in the non-effort making segment of the market.

Proposition 12. *The introduction of heterogeneity in bank screening efficiency ($\phi^b > 0$) in a sector with two or more banks compared to the competitive equilibrium without bank heterogeneity is such that it distorts effort making incentives towards less productive projects (i) when bank heterogeneity in monitoring efficiency between the most efficient bank and the other one(s) is high and (ii) when the market power of the less inefficient bank(s) is low.*

Proposition 13. *The introduction of heterogeneity in bank screening efficiency ($\phi^b > 0$) in a sector with two or more banks compared to the competitive equilibrium without bank heterogeneity is such that it reallocates resources away from entrepreneurs towards banks by (i) deteriorating the financing conditions of effort (high yield) making entrepreneurs and (ii) deteriorating the financing conditions of non-effort (medium yield) making entrepreneurs when the markup is endogenously given by the profits preserving constraint in the duopoly setting.*

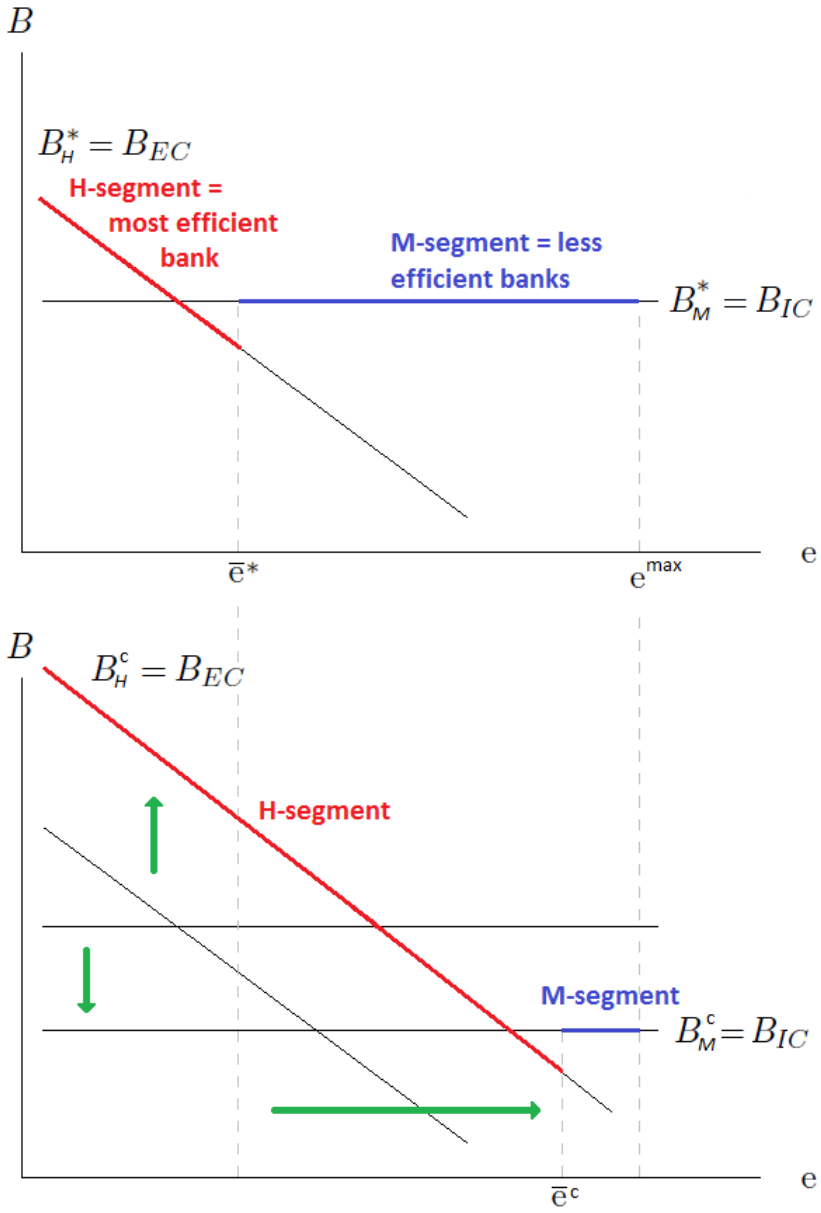
Proposition 14. *The introduction of heterogeneity in bank screening efficiency ($\phi^b > 0$) in a sector with two or more banks compared to the social optimum is such that it distorts effort making on behalf of entrepreneurs.*

Proposition 15. *The introduction of heterogeneity in bank screening efficiency ($\phi^b > 0$) in a sector with two or more banks compared to the social optimum is such that it reallocates resources away from effort making entrepreneurs; (i) it deteriorates the financing conditions of effort making entrepreneurs but (ii) may improve financing conditions of the non-effort making entrepreneurs if the less efficient banks face a low enough market power.*

So the presence of less efficient banks operating on the non-effort compatible segment of the market increases moral hazard (the extensive margin) as well as favours the missallocation of resources (the intensive margin).

Figure 8 pictures the two margins of the model with the two equilibrium, H and M-type, respectively for efficient and less efficient banks. The lower graph shows the model when there

Figure 8: Model with (up) and without (down) banking heterogeneity



is no banking heterogeneity, that is to say with no difference in the cost of monitoring; the arrow on the e axis represents the extensive margin of introducing inefficient banks: without bank heterogeneity, adverse selection is reduced and more entrepreneurs choose effort-inducing contracts. The arrows on the B axis represent the intensive margin of introducing inefficient banks: without bank heterogeneity, credits are concentrated on more productive firms.

8.2 Productivity fluctuations and banking heterogeneity

8.2.1 Moral hazard as the standard source of lending fluctuations

The competitive benchmark without banking heterogeneity already pictures the traditional source of procyclicality explored by Holmstrom and Tirole (1997) due to the fact that moral hazard is less severe when the productivity increases because it becomes relatively less attractive to shirk as the return of not diverting funds or making effort increases faster.

Proposition 16. *At the individual level, when a positive productivity shock hits, whether or not the economy features bank heterogeneity, loan contracts for high or medium yield projects become more attractive.*

8.2.2 Heterogeneous aggregate lending fluctuations

When one departs from the traditional benchmark which assumes a single positive NPV project, the model features heterogeneous lending fluctuations which are specific to each project, and this specificity spills over to the banks that finance each sector; introducing an inefficient monitoring externality creates a heterogeneous response of aggregate lending after a productivity shock. This is consistent with the fact that all banks do not necessarily modify their lending volume at the same pace.

Proposition 17. *At the individual level, when a positive productivity shock hits, whether or not the economy features bank heterogeneity :*

- *due to adverse selection, the mass of firms undertaking H-projects increases while the mass of firms undertaking M-projects decreases;*
- *due to moral hazard, M-yield loan contracts improve less than H-yield loan contracts, under the sufficient condition that the market power ψ on the M-segment of the market be large enough.*

Corollary 2. *At the aggregate level, when a productivity shock hits, whether or not the economy features bank heterogeneity, (i) lending to effort making –high yield– entrepreneurs increases while (ii) lending to low effort making –medium yield– entrepreneurs increases less, under the sufficient condition that the markup on the M-segment of the market be large enough.*

Corollary 3. *If the banking sector is heterogeneous in its screening abilities, the less efficient banks are the ones whose aggregate lending conditions fluctuate less as a result of productivity*

shocks.

Corollary 3 simply re-states corollary 2 to say that heterogeneous lending fluctuations are not driven directly by the heterogeneity in the banking industry, but rather reflect the endogenous market segmentation in the banking industry which is derived from the heterogeneity in the productive sector.

Two effects are combined in order to obtain the aggregate cyclical pattern of efficient versus less efficient bank lending; due to moral hazard, efficient bank loans are more reactive to a productivity shock because the amount lent to the entrepreneur needs to implement adequate incentives : with a lower productivity, effort making entrepreneurs have relatively more to lose than non-effort making ones, such that the efficient bank will decrease its loan size relatively faster than the less efficient banks to make sure the incentive constraint is still met. Then with a lower productivity, adverse selection will worsen such that the attractiveness of the effort-making contract for some entrepreneur will decrease since they have relatively less to gain by paying the cost of effort. So the marginal entrepreneur will switch from efficient to less efficient bank contracts as he will be now better off decreasing its effort intensity; the mass of entrepreneurs asking for efficient bank loans decreases in case of bad productivity shock, and each efficient bank loan contract reacts more to the shock, so that the two effects go in the same direction: aggregate lending by the most efficient bank decreases when the economy is hit by a negative productivity shock, i.e. it is procyclical. Conversely, for less efficient banks, the effect is ambiguous, with more entrepreneurs choosing less efficient bank loans in bad times, but still a lower loan offer per individual. Less efficient bank loans are then less responsive to productivity shocks than efficient bank ones, but may be either acyclical or countercyclical depending on the specificities of the banking sector, especially the markup on the M-yield segment of the market.

Conjecture 1. *Banks featuring a lower cyclicality of their aggregate lending may not be desirable as it may signal a lower efficiency in their monitoring abilities, and thus a specialisation in loans to less-able entrepreneurs.*

8.2.3 Bank heterogeneity as an additional source of lending fluctuation

The model presented here includes an additional source of fluctuation after a productivity shock because of the presence of banking heterogeneity.

Corollary 4. *At the individual level, when a positive productivity shock hits, the additional effect of increasing banking heterogeneity is such that the mass of firms undertaking High yield projects increases relatively more rapidly due to adverse selection.*

Indeed, the first order effect is such that the markup on the H-segment of the market $\frac{1}{Q^{**}}$ increases when banking heterogeneity increases, that is to say when the efficiency gap in monitoring technologies increases compared to the most efficient bank. But the markup decreases when productivity increases, because the intensity of the competition between banks increases in order to finance increasingly profitable H projects relative to M-yield ones.

Now the second order effect is such that this markup reduces relatively faster when the banking heterogeneity marginally increases after a positive productivity shock. Indeed more banking heterogeneity means less competition, more rent extraction and consequently lower loan size asked by entrepreneurs. Thus the marginal profit of the efficient bank it retrieves from a positive productivity shock decreases as entrepreneurs ask for relatively less loans. Henceforth, the difference of profitability between the most efficient bank and the less efficient competitors is marginally smaller, which is equivalent to a tightening of the competition and a reduction of the markup at the margin, even if the latter increases in absolute amounts.

The above proposition states that more banking heterogeneity marginally increases the fluctuations of individual lending to effort making entrepreneurs after a positive productivity shock. But banking heterogeneity has an ambiguous impact on the fluctuation of average productivity following a positive productivity shock, as it indeed increases relatively faster the pool of effort making entrepreneurs undertaking H-yield projects, but increasing banking heterogeneity also exacerbates adverse selection and reduces the pool of effort making entrepreneurs.

Corollary 5. *When a positive productivity shock hits, the additional effect of increasing banking heterogeneity on average productivity fluctuation is ambiguous.*

Proposition 18. *When a productivity shock hits, in an economy with several heterogeneous banks, the pool of effort-making entrepreneurs is less volatile compared to the social optimum, provided that the private benefit of not behaving is not too high.*

The reason for the restriction of proposition 18 goes as follows; if the private benefit was too high, the less efficient banks on the M-segment of the market would have to offer relatively less attractive loans when a negative shock hits, because otherwise the entrepreneur would be better off diverting funds from a large loan when the return on behaving is low. Then as a

result, if private benefits were excessively high, relatively less entrepreneurs would be attracted by Medium-yield contracts whose conditions are deteriorating faster; so, compared to the social optimum, below a certain threshold of private benefits, this effect is dominated for sure, which means a lower volatility of the pool of effort-making entrepreneurs.

9 Numerical Simulation

I now provide some numerical simulations of the general model of n banks for different intensities of the banking heterogeneity in monitoring efficiency and different market structures associated with more or less competition. Table 1 shows the specific parametrisation used thereafter, which satisfies assumptions 1 to 8.

9.1 Mechanics of the model

Graph 9 and 10 describes the mechanics inherent to the model. The vertical lines represent the threshold entrepreneur which is indifferent between undertaking a H-yield project while paying the cost of effort or undertaking a lower M-yield project where the effort cost is reduced. Thus this line separates the market into two components; to the left of the distribution, all entrepreneurs with a small cost of effort ask a loan to the most efficient bank, and to the right of the line, entrepreneurs with a larger cost of effort are not incentivised to pick the largest positive NPV project and obtain a loan from the less efficient banks.

Note that the competitive equilibrium without banking heterogeneity is already a second best and fails to maximise aggregate surplus due to the existence of an externality: the banks do not internalise the implication of bank competition in contracts on the decision of the entrepreneurs who will have different incentives to undertake effort or not depending on the set of contracts being offered to them.

One can readily see that the introduction of banking heterogeneity introduces an additional wedge both in terms of allocative and incentive distortions, which can be understood as an additional externality created by the existence of differences in monitoring abilities, which is in fact a positive externality for the most efficient bank as it allows it to extract a rent from its dominant position, but a negative externality at the social level as banks fail to factor in the effort incentives implications. Compared to the banking system without banking heterogeneity or to the optimal benchmark, on the one hand, the loan size offered to High yield entrepreneurs

is lower after the introduction of banking heterogeneity, and on the other hand the number of entrepreneurs deciding to provide sufficient effort to undertake the project with the largest social net present value is reduced and the threshold entrepreneur moves to the left of the distribution.

9.2 Allocative distortions and bank-level fluctuations

Then the next set of pictures shows the evolution of the two types of distortions as a function of the productivity parameter and the magnitude of the fluctuation when banking heterogeneity increases.

Graph 11 pictures the *extensive margin* which comes from the adverse selection problem. When a productivity shock hits, the mass of entrepreneurs evolves along the e axis: adverse selection will worsen in time of crisis, such that the fringe of entrepreneurs in the middle of the distribution will find it less profitable to behave and may no longer be willing to pay the cost of effort relative to lower productivity levels. Henceforth, some entrepreneurs will find it more profitable to deviate and relax the pressure of the monitoring by switching to effort-free packages offered by less efficient banks, that is to say switching from the red area to the blue+cyan one, where the cyan area corresponds to the range of entrepreneurs which would have chosen to undertake effort at the optimum. So, in the terminology of macroeconomists, *the mass of firms financed by less efficient banks is countercyclical.* In addition, it can be readily seen by observing the slopes of the threshold entrepreneur that an increase in banking heterogeneity increases the fluctuations in the incentives to make effort as a result of a shock on productivity.

In addition, graphs 12 and 13 picture the *intensive margin* respectively on the H and M-segment of the market, which arises as a result of the moral hazard present in the model; when a productivity shock hits, the size of the loan offered to effort-making entrepreneur will decrease faster since the effort compatibility constraint will be tightening faster as effort-making entrepreneurs have relatively more to lose ($p_H\alpha$ versus $p_M\alpha$). So *loan size offered by less efficient banks will be less procyclical than the one of efficient banks*, which can be readily seen on the graph: the slope of loan size provided by the efficient bank on the H segment of the market is larger than the slope of loan size provided by less efficient banks on the M segment of the market.

9.3 Aggregate fluctuations of productivity and lending

I now turn to the description of the implication of the model on productivity levels and aggregate fluctuations when banking heterogeneity increases and for different intensities of the markup in the banking industry in the M-segment of the market [7].

Pictures 14 and 15 show the impact of banking heterogeneity on respectively average and weighted average productivity. Note that realised productivity is different from the common productivity parameter α, as each entrepreneur, given the productivity level common to the whole economy, can decide to choose a higher or lower productivity project depending on its individual return of providing effort. When banking heterogeneity increases, due to the rent extraction and the resulting adverse selection, less entrepreneurs are ready to make efforts so that average realised productivity across entrepreneurs decreases, that is to say the productivity of an average entrepreneur in the economy decreases. Likewise, as the surplus is captured away from entrepreneurs, only a smaller loan size can sustain effort compatibility because of the presence of moral hazard; so the two effects are combined and more banking heterogeneity leads to lower realised weighted average productivity, that is to say the average productivity level observed in the economy while weighting by projects size of each entrepreneur.

Now when the markup in the M-segment of the market increases, that is to say the competition for the financing of the Medium-yield projects is not perfect and less efficient banks still extract a rent, the average or aggregate productivity increases despite the presence of banking heterogeneity. This arises as it reduces the relative gap between H or M-compatible contracts: the M-contract becomes less attractive so that entrepreneurs have less incentives to pick a Medium yield project as the opportunity cost of making effort decreases. Thus some entrepreneurs will be better off undertaking H-yield projects despite the presence of the endogenous markup introduced by banking heterogeneity: a markup in the main segment of the market can get the extensive margin right so that more entrepreneurs choose to undertake high yield projects in the niche, but at the cost of a depressed intensive margin with additional rent extraction. Thus a high enough markup in the market for medium yield projects can push the realised weighted average productivity of the economy close to the optimum, which is but detrimental to total production.

Then figure 16 pictures the aggregate lending of the whole economy, after aggregating the

7. The markup in the M-segment of the market is derived endogenously only in the specific duopoly case, while the markup on the H-segment of the market –the most important one corresponding to effort decisions– is always endogenously pinned down.

two segments of the market, that is to say taking the integral of bank loans over the whole range of entrepreneurs. It clearly shows that introducing banking heterogeneity, via both the intensive and extensive margins discussed above, leads to a lower level of aggregate lending compared to the situation without heterogeneity in banking efficiency. Note that aggregate lending could be larger in the competitive case as opposed to the optimal situation, as aggregate lending volumes is not sufficient to ensure allocative efficiency across entrepreneurs willing or not to make effort.

If competition is not perfect in the M-segment of the market for less productive projects, aggregate lending volumes further decrease; indeed, more entrepreneurs undertake High yield projects, but the size of loans to each non-effort making entrepreneurs is reduced as their smaller surplus cannot prevent them from deviating and shirking if the loan size they get is not reduced.

Moreover, whether the endogenous markup on the H-segment of the market or the exogenous (endogenous if n=2) markup on the M-segment increases, the response of aggregate lending to a productivity shock decreases and gets "less procyclical". Thus this confirms conjecture 1 stated at the bank level: a lower cyclicality of aggregate bank lending may signal a larger heterogeneity in the efficiency of bank monitoring abilities, which is shown to be detrimental to average productivity; as attractive a lower volatility might seem, it may be the counterpart of allocative and incentives distortions.

Conjecture 2. *A banking system featuring a lower cyclicality of aggregate lending might not be desirable as it may signal allocative and incentives distortions steaming from banking heterogeneity in bank monitoring efficiencies.*

Last, figure 17 displays the total output produced in each case. As expected, the case of fully homogeneous and thus competitive banking is already a second best compared to the optimum where all entrepreneurs that can make effort use their full potential, while introducing banking heterogeneity reduces total output, which is further reduced when the competition between banks on the main segment of the market becomes oligopolistic.

10 Conclusion

This paper is a first attempt to connect heterogeneity in bank efficiency with lending fluctuations and allocative efficiency in an endogenously fragmented market for loans. The presence of banking heterogeneity in monitoring costs highlights a trade-off between a lower reaction of aggregate lending to productivity shocks and a larger missallocation of ressources towards a

growing pool of less productive entrepreneurs. In essence, introducing a wedge in monitoring efficiencies on the market that requires monitoring is akin to an externality as it spills over to markets which do not require monitoring, and thus modify bank loan contracts even when monitoring is not required. When the banking heterogeneity in monitoring efficiency increases, an endogenous niche arises as a result of moral hazard and adverse selection, which can explain the observed heterogeneity in individual bank lending fluctuations. Moreover this endogenous market segmentation is directly connected to the introduction of two positive NPV projects, which departs form standard corporate finance theory that would otherwise overlook the effect of banking heterogeneity and market fragmentation. Thus lending fluctuations on the different segments of the market become heterogeneous and the rent extraction differential across market segments turns out to be detrimental to allocation efficiency.

The next step to study the question at hand, namely the effect of monitoring heterogeneity on both the cyclical dimension and cross-sectional allocation, would be to integrate the one period model into a multi-period general equilibrium model ; such a framework would be able to go beyond the mere comparative statics of the effect of productivity shocks, in order to analyse the effect of heterogeneity on the whole business cycle. Then successive periods would allow for proper lending relationship, i.e. the introduction of a cost of switching over time from a bank in a niche to another bank loan contract; this effect is likely to dampen the impact of banking heterogeneity on lending fluctuations. In addition, adding entrepreneurs' entry to the model would require to add some heterogeneity among the medium-yield entrepreneurs –which are ex-post identical–, so that some of them would prefer not to enter the productive sector. But the analysis would likely go through, depending partly on the relative magnitude of the switching and exiting decisions.

Taking stock of the model and its simulations, one should be cautious when observing less cyclical lending fluctuations, both at the bank or country-wide level. Here, heterogeneity in bank lending fluctuations reflects the heterogeneity on the asset side with projects of both variable profitability and variable costs ; then heterogeneity in bank monitoring efficiency explains why certain banks would specialise in some segment of the market rather than serve the whole market. Henceforth, banks featuring a lower cyclicality of their aggregate lending may not be desirable as it may signal a lower efficiency in their monitoring abilities. Likewise, a banking system featuring a lower cyclicality of aggregate lending might not be desirable as it may signal allocative and incentives distortions steaming from banking heterogeneity in monitoring

efficiencies. As a result, the presence of heterogeneity decreases average productivity as it increases adverse selection by entrepreneurs as well as favours rent extractions by banks.

In terms of policy implications when implementing structural reforms of a banking sector, three pitfalls should be avoided; first, the model that replicates new stylised facts on banking heterogeneity calls for more caution in terms of supporting structures associated with more stable aggregate lending as it may come as a result of a monitoring inefficiency externality. Second, when one carries out analyses aiming at investigating the efficiency of the banking sector to channel resources to the most promising projects, one should consider not only overall banking efficiency but also look at the cross-sectional distribution of banks as average banking efficiency is not enough to ensure productive efficiency. Last, reforms implementing new regulations which would have an impact on monitoring or screening incentives (such as capital or liquidity regulation) should also be careful not to increase banking heterogeneity that would create possibilities of rent extractions by some institutions and distort the adverse selection process by borrowers.

References

Adrian, T. and H. S. Shin (2010) "Liquidity and Leverage," *Journal of Financial Intermediation*, Vol. 19, pp. 418–437. 1

Alessi, L. and C. Detken (2009) "Real time early warning indicators for costly asset price boom/bust cycles - a role for global liquidity," ECB Working Paper Series, 1039. 1

Bertay, A. C., A. Demirgüç-Kunt, and H. Huizinga (2012) "Bank Ownership and Credit over the Business Cycle: Is Lending by State Banks Less Procyclical?", Working paper, World Bank, 6110. 1, 3

Demirgüç-Kunt, A., T. Beck, and P. Honohan (2008) "Finance for all?: Policies and pitfalls in expanding access," Tilburg University Working Paper. 1

Duprey, T. (2012) "Bank Ownership and Credit Cycle : the lower sensitivity of public bank lending to the business cycle," Working paper 411, Bank of France. 1, 3

Duprey, T. and M. Lé (2012) "Bankscope dataset: getting started," working paper. 3

Giannone, D., M. Lenza, and L. Reichlin (2011) "Market Freedom and the Global Recession," *IMF Economic Review*, Vol. 59, pp. 111–135. 1, 3

Hartmann, P., F. Heider, M. Lo Duca, and E. Pappaioannou (2007) "The role of financial markets and innovation for productivity and growth in Europe," ECB Occasional Paper, 72. 1

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. 1, 3, 8.2.1

Kaminsky, G. and K. Reinhart (1999) "The twin crises: The causes of banking and balance-of-payments problems," *American Economic Review*, Vol. 89, pp. 473–500. 1

King, R. and R. Levine (1993) "Finance and growth: Schumpeter might be right," *Quarterly Journal of Economics*, Vol. 108, pp. 717–737. 1

Kose, A., E. Prasad, K. Rogoff, and S-J. Wei (2003) "Effects of financial globalization on developing countries: Some empirical evidence," International Monetary Fund Occasional Paper, 220. 1

La Porta, R., F. Lopez-De-Silanes, and A. Shleifer (2002) "Government Ownership of Banks," *The Journal of Finance*, Vol. 57, No. 1, pp. 265–301. 3

Levchenko, A., R. Ranciere, and M. Thoenig (2009) "Growth and risk at the industry level: The real effects of financial liberalization," *Journal of Development Economics*, Vol. 89, pp. 210–222. 1

Levine, R. (2005) "Finance and growth: Theory, evidence, and mechanisms," in P. Aghion and S. Durlauf eds. *The Handbook of Economic Growth*, Amsterdam: North-Holland. 1

Micco, A. and U. Panizza (2006) "Bank ownership and lending behavior," *Economics Letters*, Vol. 93, pp. 248–254. 1, 3

Rajan, R. and L. Zingales (1998) "Financial dependence and growth," *American Economic Review*, Vol. 88, pp. 559–586. 1

Reichlin, P. (2004) *Credit, Intermediation and the Macroeconomy - Readings and Perspectives*, Chap. Credit markets and the Macroeconomy, Modern Financial Theory: Oxford University Press. 3

Wurgler, J. (2000) "Financial markets and the allocation of capital," *Journal of Financial Economics*, Vol. 58, pp. 187–214. 1

A Proofs

Proof of Proposition 3. The threshold entrepreneur $\bar{e}^c$ above which entrepreneurs choose to undertake the M-yield project is given by :

$$\begin{aligned}
C_M^c(R_M^c, B_M^c) &\succ C_H^c(R_H^c, B_H^c) \\
p_M\left(\alpha(B_M^c + A) - R_M^c B_M^c\right) &\geq p_H\left(\alpha(B_H^c + A) - R_H^c B_H^c\right) - ek_H^c \\
p_M \alpha \frac{R_M^c}{\frac{b}{\Delta_M} + R_M^c - \alpha} A - p_M R_M^c \frac{\alpha - \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} A &\geq p_H\left(\alpha(B_H^c + A) - R_H^c B_H^c\right) - ek_H^c \\
\frac{p_M R_M^c \frac{b}{\Delta_M}}{(B_H^c + A)(\frac{b}{\Delta_M} + R_M^c - \alpha)} A &\geq p_H \alpha - p_H \frac{R_H^c B_H^c}{B_H^c + A} - e \\
\frac{p_M R_M^c \frac{b}{\Delta_M}}{(B_H^c + A)(\frac{b}{\Delta_M} + R_M^c - \alpha)} A &\geq p_H \alpha - p_H B_H^c \frac{\frac{e}{\Delta_H} + R_H^c - \alpha}{A} - e \\
\frac{p_M R_M^c \frac{b}{\Delta_M}}{(B_H^c + A)(\frac{b}{\Delta_M} + R_M^c - \alpha)} A &\geq p_H \alpha - p_H(\alpha - \frac{e}{\Delta_H}) - e \\
\frac{p_M R_M^c \frac{b}{\Delta_M}}{(B_H^c + A)(\frac{b}{\Delta_M} + R_M^c - \alpha)} A &\geq e\frac{p_M}{\Delta_H} \\
\frac{(\frac{e}{\Delta_H} + R_H^c - \alpha)}{R_H^c} \frac{p_M R_M^c \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} &\geq e\frac{p_M}{\Delta_H} \\
\left(\frac{(\frac{e}{\Delta_H} - \alpha)}{R_H^c} + 1\right) \frac{p_M R_M^c \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} &\geq e\frac{p_M}{\Delta_H} \\
\left(\frac{(\frac{e}{\Delta_H} - \alpha)(p_H - \frac{c\Delta_H}{\alpha\Delta_H - e})}{r} + 1\right) \frac{p_M R_M^c \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} &\geq e\frac{p_M}{\Delta_H} \\
\left(\frac{ep_H - \alpha\Delta_H p_H + c\Delta_H}{r\Delta_H} + 1\right) \frac{p_M R_M^c \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} &\geq e\frac{p_M}{\Delta_H} \\
e\left(\frac{p_H}{r\Delta_H} \frac{p_M R_M^c \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} - \frac{p_M}{\Delta_H}\right) &\geq \frac{\left(\frac{\alpha p_H - c}{r} + 1\right) p_M R_M^c \frac{b}{\Delta_M}}{\frac{b}{\Delta_M} + R_M^c - \alpha} \\
\frac{e}{\Delta_H}\left(\frac{p_H p_M R_M^c b}{r\Delta_M} - p_M(\frac{b}{\Delta_M} + R_M^c - \alpha)\right) &\geq \frac{p_M R_M^c b}{r\Delta_M}(\alpha p_H - c - r) \\
e &\geq \Delta_H \frac{\alpha p_H - c - r}{p_H - \frac{r\Delta_M}{b} - \frac{r}{R_M^c}(1 - \frac{\Delta_M \alpha}{b})}
\end{aligned} \tag{28}$$

Replacing the interest rate $R_M^c = \frac{r}{p_M}$ in equation (28) –which gives the expression for the threshold entrepreneur being indifferent between the H or M-segment of the market:

$$\begin{aligned}
e &\geq \frac{\frac{b\Delta_H}{\Delta_M}(\alpha p_H - c - r)}{\frac{p_H b}{\Delta_M} - p_M(\frac{b}{\Delta_M} + \frac{r}{p_M} - \alpha)} \\
e &\geq \frac{\Delta_H(\alpha p_H - c - r)}{p_H - \frac{p_M \Delta_M}{b}(\frac{b}{\Delta_M} + \frac{r}{p_M} - \alpha)} \\
e &\geq \frac{\Delta_H(\alpha p_H - c - r)}{\Delta_H - r\frac{\Delta_M}{b} + \alpha p_M \frac{\Delta_M}{b}} \equiv \bar{e}^c
\end{aligned}$$

So in the fully competitive case, entrepreneurs below the threshold $\bar{e}^c$ will choose the H-contract which implements effort. □

Proof of Proposition 4. The contract that maximises the surplus of entrepreneurs on the H-segment of the market is the same as the one showed in Proposition 1; likewise the size of the loan financing M-yield projects is given as in proposition 2. But now, in order to make sure all

entrepreneurs who can undertake effort are properly incentivised, the M-contract is designed such that the marginal entrepreneur able to make effort $\bar{e}^o \equiv \alpha\Delta_H - c > \bar{e}^c$ finds it profitable to choose the H-contract which is effort compatible. Since $B_M^o = \bar{B}_{IC}$, R_M^o is chosen such that for entrepreneur $\bar{e}^o$:

$$\begin{aligned} C_H^o(R_H^o, B_H^o) &\succ C_M^o(R_M^o, B_M^o) \\ p_H\left(\alpha(B_H^o + A) - R_H^o B_H^o\right) - \bar{e}^o k_H^o &\geq p_M\left(\alpha(B_M^o + A) - R_M^o B_M^o\right) \\ R_M^o &= \frac{p_M\alpha + \frac{(p_M\alpha - r)\frac{p_M}{\Delta_H}c}{r - \frac{p_M}{\Delta_H}c}}{p_M + \frac{(p_M\alpha - r)\frac{p_M}{\Delta_H}c}{(r - \frac{p_M}{\Delta_H}c)(\alpha - \frac{b}{\Delta_M})}} \end{aligned}$$

□

Proof of Proposition 6. The loan size B_M^* is obtained as in proposition 2 in the absence of bank heterogeneity since the profits of individual entrepreneurs is increasing in loan size and $\bar{B}_{IC}$ is the maximal amount which can sustain a positive NPV project.

The interest rate R_M^* is obtained as follows. First recognize that the bank with high abilities and lower monitoring costs has an incentive not to enter the M-segment of the market as it is profits improving for the rent it derives from the H-yield segment of the market. The marginal entrepreneur indifferent between H or M-types of contracts is derived as in proposition 3 such that:

$$\begin{aligned} C_M^h(R, B_M^*) &\succ C_H^l(R_H^*, B_H^*) \\ p_M\left(\alpha(B_M^* + A) - RB_M^*\right) &\geq p_H\left(\alpha(B_H^* + A) - R_H^* B_H^*\right) - ek_H^* \\ e &\geq \Delta_H \frac{\alpha p_H - c^h - r}{p_H - \frac{r\Delta_M}{b} - \frac{r}{R}(1 - \frac{\Delta_M\alpha}{b})} \equiv \bar{e}^* \end{aligned} \tag{29}$$

The efficient bank may now decide not to make M-loans if by doing so it is able to attract more entrepreneurs willing to enter an H-contract, that is to say if some entrepreneurs close to the threshold decide to switch from M-yield to H-yield projects when the market power on the M-segment of the market increases so that the interest rate R_M^* increases above its competitive

value $\frac{r}{p_M}$. That is to say one must have:

$$\begin{aligned} \frac{\partial \bar{e}^*}{\partial R} &> 0 \\ \Delta_H(\alpha p_H - c^{pu} - r)\frac{\frac{r}{R^2}\frac{\Delta_M\alpha - b}{b}}{(p_H - \frac{r\Delta_M}{b} - \frac{r}{R}(1 - \frac{\Delta_M\alpha}{b}))^2} &> 0 \end{aligned}$$

which is indeed positive from assumption 4, so that the more efficient bank indeed has an incentive to allow for a rent to be extracted by the inefficient bank with an offered interest rate $R_M^* = \frac{r}{p_M} + \psi$ above the competitive one, with ψ being the intensity of the market power to be determined next.

Second, once the less efficient bank is left on the M-segment of the market with some market power, it will decide to maximise its now positive profits over the whole range of entrepreneurs choosing M-yield projects.

$$\max_{R>r/p_M} \Pi^h_{total} \leftrightarrow \max_{R>r/p_M} \left(e^{max} - \bar{e}^*\right)\left(p_M R - r\right) B_M^* \tag{30}$$

The first derivative is given by:

$$\frac{\partial \Pi^h_{total}}{\partial R} = -\frac{\bar{e}^*}{\partial R}\left(p_M R - r\right) B_M^* + \left(e^{max} - \bar{e}^*\right)\left(p_M B_M^* + \left(p_M R - r\right)\frac{\partial B_M^*}{\partial R}\right)$$

In order to show that there indeed exist a maximum profit for R_M^*, I first show that starting at the competitive equilibrium where $\psi = 0$, marginal profits are positive if the interest rate was to increase, i.e. the market power ψ increases. Then I show that for $\psi \to \infty$, marginal profits tend to zero from negative values. Thus if profits starting at zero initially increase and then decrease, provided the variables are continuous, there indeed exists a positive markup which maximises aggregate profits of the less efficient bank for an interest rate $R_M^* = r/p_M + psi^* > r/p_M$.

– for $\psi = 0$, $p_M R = r$ so that I indeed have:

$$\frac{\partial \Pi^h_{total}}{\partial R} = \left(e^{max} - \bar{e}^*\right) p_M B_M^* > 0$$

– for $\psi \to \infty$, notice that the first term of $\frac{\partial \Pi^h_{total}}{\partial R}$ tends to 0 by negative values since:

$$\begin{aligned} \lim_{\psi\to\infty} \frac{\bar{e}^*}{\partial R} &= 0_+ \\ \lim_{\psi\to\infty} (p_M R - r) B^*_M &= \frac{\alpha - b/\Delta_M}{\frac{1}{p_M - r/R} - \frac{\alpha - b/\Delta_M}{p_M R - r}} A \\ &= p_M(\alpha - \frac{b}{\Delta_M})A \end{aligned}$$

then for the second term of $\frac{\partial \Pi^h_{total}}{\partial R}$:

$$\lim_{\psi\to\infty} (e^{max} - \bar{e}^*) = e^{max} - \Delta_H \frac{\alpha p_H - c^h - r}{p_H - \frac{r\Delta_M}{b}} < 0$$

which is negative for e^{max} sufficiently small, and :

$$\begin{aligned} \lim_{\psi\to\infty} \left(p_M B^*_M + (p_M R - r)\frac{\partial B^*_M}{\partial R} \right) &= p_M \frac{\alpha - \frac{b}{\Delta_M}}{R - (\alpha - \frac{b}{\Delta_M})} A - \frac{(p_M R - r)\left(\alpha - \frac{b}{\Delta_M}\right)}{\left(R - (\alpha - \frac{b}{\Delta_M})\right)^2} A \\ &= \lim_{\psi\to\infty} \left(\alpha - \frac{b}{\Delta_M}\right) A \frac{r - p_M\left(\alpha - \frac{b}{\Delta_M}\right)}{R - \left(\alpha - \frac{b}{\Delta_M}\right)^2} \\ &= 0_+ \end{aligned}$$

from assumption 4 and provided that the NPV of the M project is not too high $p_M\alpha - r < p_M \frac{b}{\Delta_M}$.

□

Proof of Proposition 7. The cutoff entrepreneur $\bar{e}^*$, given in the proof of proposition 6, is such that $e > \bar{e}^* \Leftrightarrow C^{*h}_M(R^*_M, B^*_M) \succ C^{*l}_H(R^*_H, B^*_H)$ where $\{h, l\}$ represent respectively the high monitoring cost or inefficient bank and the low monitoring cost or efficient bank.

□

Proof of Proposition 10.

$$\begin{aligned} C_M^h(R_M^{**}, B_M^{**}) &\succ C_H^l(R_H^{**}, B_H^*) \\ p_M\left(\alpha(B_M^{**}+A) - R_M^{**}B_M^{**}\right) &\geq p_H\left(\alpha(B_H^{**}+A) - R_H^{**}B_H^{**}\right) - ek_H^{**} \\ e &\geq \Delta_H \frac{\alpha p_H - c^l(1+\phi^{b=2}) - r}{p_H - \frac{r\Delta_M}{b} - \frac{r}{R_M^{**}}(1-\frac{\Delta_M\alpha}{b})} \\ e &\geq \Delta_H \frac{\alpha p_H - c^l(1+\phi^{b=2}) - r}{p_H - \frac{r\Delta_M}{b} + \frac{r}{\frac{r}{p_M}+\psi}(\frac{\Delta_M\alpha}{b}-1)} \equiv \bar{e}^{**} \end{aligned}$$

□

Proof of Proposition 11. The intensity of bank inefficiency in monitoring is measured by ϕ:

- in the duopoly case, $c^h = c^l(1+\phi)$ with $\phi > 0$; it follows from proposition 7 that $\frac{\partial \bar{e}^*}{\partial \phi} < 0$
- in the case with n banks, $c^{b=2} = c^l(1+\phi^{b=2})$ with $c^b \in \left\{c^l, c^{b=2}, ..., c^{b=n}\right\}$ and $\phi^b \in \left\{0, \phi^{b=2}, ..., \phi^{b=n}\right\}$; it follows from proposition 10 that $\frac{\partial \bar{e}^{**}}{\partial \phi} < 0$

so the larger the inefficiency, the smaller the range of effort making entrepreneurs given by $\bar{e}^* - e_{min}$ or $\bar{e}^{**} - e_{min}$. So more heterogeneity in screening efficiency between the first two banks ranked by monitoring efficiency means stronger adverse selection towards less productive projects. □

Proof of Corollary 1. Average productivity varies when the intensity of the heterogeneity in the banking sector ϕ increases. In the duopoly case:

$$\frac{\partial}{\partial \phi}\left(\alpha \frac{p_H * (\bar{e}^* - e^{min}) + p_M * (e^{max} - \bar{e}^*)}{e^{max} - e^{min}}\right) = \frac{\alpha}{e^{max} - e^{min}}(p_H - p_M)\frac{\partial \bar{e}^*}{\partial \phi} < 0 \tag{31}$$

where it follows from proposition 11 that $\frac{\partial \bar{e}^*}{\partial \phi} < 0$. Likewise for $\bar{e}^{**}$ in the case of n heterogeneous banks. □

Proof of Proposition 12. Banking heterogeneity distorts effort making incentives in the following two cases:

- It follows from propositions 3 and 7 that bank heterogeneity in bank efficiency distorts incentives and increases adverse selection towards non-effort compatible contracts if I have

$\bar{e}^c > \bar{e}^*$:

$$\begin{aligned}
\bar{e}^c &> \bar{e}^* \\
\frac{\Delta_H(\alpha p_H - c^l - r)}{\Delta_H - r\frac{\Delta_M}{b} + \alpha p_M \frac{\Delta_M}{b}} &> \Delta_H \frac{\alpha p_H - c^l(1+\phi) - r}{p_H - \frac{r\Delta_M}{b} - \frac{r}{R_M^*}(1 - \frac{\Delta_M \alpha}{b})} \\
(\alpha p_H - c^l - r)(-\frac{r}{R_M^*}(1 - \frac{\Delta_M \alpha}{b}) + p_M - p_M \frac{\alpha \Delta_M}{b}) &> -c^l \phi(\Delta_H - r\frac{\Delta_M}{b} + \alpha p_M \frac{\Delta_M}{b}) \\
(\alpha p_H - c^l - r)(p_M - \frac{r}{R_M^*})(1 - \frac{\Delta_M \alpha}{b}) &> -c^l \phi(\Delta_H - r\frac{\Delta_M}{b} + \alpha p_M \frac{\Delta_M}{b})
\end{aligned}$$

The left hand side of the last expression is negative (first term positive, second positive since $R_M^* > \frac{r}{p_M}$, and third negative from assumption 4) but the right hand side is negative as well. In fact two effects work in opposite directions. By introducing an inefficient bank, you decrease competition on the H-segment of the market and create a rent for the efficient bank on this segment, which should lower incentives to undertake efforts. In the meantime, this profits-preserving behaviour for the efficient bank increases the market power of the inefficient bank on the M-segment of the market, which should result in lower incentives to shirk and de facto more incentives to turn to effort making contracts.

Recall that the cost of monitoring of the less efficient bank is given by $c^h = c^l(1+\phi)$ such that ϕ is an index of bank monitoring efficiency when there is bank heterogeneity. If ϕ is too high, then more banking heterogeneity will allow for an excessively high rent to be extracted on the H-segment of the market which will discourage effort so that indeed $\bar{e}^c > \bar{e}^*$. The same analysis applies in the case of n heterogeneous banks, with $\bar{e}^c > \bar{e}^{**}$.

– Recall that $R_M^* = \frac{r}{p_M} + \psi$ with ψ the intensity of the market power on the M-segment of the market which is endogenously determined in proposition 6 for the duopoly case. If ψ is too small, that is to say if the financing condition of the M-contract becomes too attractive then $\bar{e}^c > \bar{e}^*$, that is too say there is an excessive number of entrepreneurs choosing not to undertake efforts in the case with bank heterogeneity.

□

Proof of Proposition 13. It follows from Proposition 3 that banks do not extract any rent in the fully competitive case and offer fair contracts.

Banking heterogeneity deteriorates the loan contracts offered to high yield entrepreneurs as banks extract a rent.

Alternatively, one can see that, for the duopoly case:

$$\frac{\partial B_H^*}{\partial \phi} = \frac{\partial B_H^*}{\partial R_H^*}\frac{\partial R_H^*}{\partial Q}\frac{\partial Q}{\partial \phi} < 0$$

Likewise for the case of n heterogeneous banks:

$$\frac{\partial B_H^{**}}{\partial \phi} = \frac{\partial B_H^{**}}{\partial R_H^{**}}\frac{\partial R_H^{**}}{\partial Q}\frac{\partial Q}{\partial \phi^{b=2}} < 0$$

– Banking heterogeneity deteriorates the loan contracts offered to medium yield entrepreneurs if it allows banks to extract a rent; in the specific duopoly case, banking heterogeneity endogenously creates a markup in the M-yield segment of the market due to the profits preserving constraint. So from the equilibrium loan size on the M-segment of the market given by the Incentive Compatibility constraint, one has:

$$\bar{B}_{IC}\left(B_M^* = \frac{r}{p_M} > \frac{r}{p_M}\right) < \bar{B}_{IC}\left(B_M^c = \frac{r}{p_M}\right) \Leftrightarrow B_M^* < B_M^c$$

But in the case with several heterogeneous banks, the equilibrium loan contract on the M-segment of the market does not deteriorate when heterogeneity increases but does deteriorate due to the presence of the markup in banking which is not expressively modeled as an endogenous outcome of banking heterogeneity:

$$\frac{\partial B_M^{**}}{\partial \phi} = \frac{\partial B_M^{**}}{\partial R_M^{**}}\frac{\partial R_M^{**}}{\partial \phi^{b=2}} = 0$$

$$\bar{B}_{IC}\left(B_M^{**} = \frac{r}{p_M + \psi} > \frac{r}{p_M}\right) < \bar{B}_{IC}\left(B_M^c = \frac{r}{p_M}\right) \Leftrightarrow B_M^{**} < B_M^c$$

□

Proof of Proposition 14. It follows from the definition of the optimum in proposition 4 that entrepreneurs above $\bar{e}^o$ could never realise a surplus sufficient enough to cover their cost of effort and compensate the bank for the monitoring cost, so that this gives an upper bound to the entrepreneurs which could decide to choose H or M. So in the case of banking heterogeneity, the efficient bank cannot expect to attract entrepreneurs above $\bar{e}^o$ by allowing for additional market power on the M-segment of the market, so that one necessarily have the following inequality $\bar{e}^* < \bar{e}^o$ whatever the intensity of market power on the M-segment of the market. The inequality is strict as the rent extraction on the H-segment due to the presence of a less

efficient bank makes the most efficient bank ask for a compensation strictly higher than its cost of monitoring, while, at the optimum $\bar{e}^o$, only a fair compensation of the monitoring cost c is considered. Henceforth, together with proposition 12, I have $\bar{e}^* \left(R_M^*\right) < \bar{e}^c \left(R_M^c\right) < \bar{e}^o \left(R_M^o\right)$.

So the range of entrepreneurs $[\bar{e}^*, \bar{e}^o]$ fails to provide effort in case of heterogeneity in bank monitoring efficiency. The same argument applies to the case of n heterogeneous banks. □

Proof of Proposition 15. From proposition 5 or 8, it is obvious that introducing bank heterogeneity within two or more banks creates a rent on the H-segment of the market so that it deteriorates the financing conditions of high yield projects undertaken by effort making entrepreneurs.

In the optimal case, the key variable is $R_M^o > \frac{r}{p_M}$ which is used to introduce a wedge high enough to discourage entrepreneurs from self-selecting M-yield projects. So depending on the intensity of the market power ψ available to the inefficient bank on the M-segment of the market, one can have $R_M^o > R_M^*$ in which case the financing conditions of non-effort making entrepreneurs is improved. The same argument applies to the case of several heterogeneous banks. □

Proof of Proposition 16. The same proof holds if I consider the fully competitive economy without bank heterogeneity, that is to say by setting $\phi^b = 0$. I consider here the economy with banking heterogeneity.

– The interest rate offered by the most efficient banks decreases when a productivity shock hits:

$$\begin{aligned} \frac{\partial R_H^{**}}{\partial \alpha} &= \frac{\partial}{\partial \alpha} \frac{r}{p_H - \frac{c^h \Delta_H}{\alpha \Delta_H - e}} \\ &= -\frac{r c^h}{(p_H - \frac{c^h \Delta_H}{\alpha \Delta_H - e})^2 (\alpha - e/\Delta_H)^2} \\ &< 0 \end{aligned}$$

which comes from the fact that the markup $\frac{1}{Q^{**}}$ is countercyclical. So individual loan size offered by the most efficient bank is positively associated with productivity shocks:

$$\frac{\partial B_H^{**}}{\partial \alpha} = \frac{\partial B_H^{**}}{\partial R_H^{**}} \frac{\partial R_H^{**}}{\partial \alpha} > 0$$

– Individual loan size offered by less efficient bank increases when a productivity shock hits:

$$\begin{aligned}\frac{\partial B_M^{**}}{\partial \alpha} &= \frac{\partial \frac{\alpha - b/\Delta_M}{(b/\Delta_M + R_M^{**} - \alpha)} A}{\partial \alpha} \\ &= \frac{(b/\Delta_M + R_M^{**} - \alpha) + (\alpha - b/\Delta_M)}{(b/\Delta_M + R_M^{**} - \alpha)^2} A \\ &= \frac{R_M^{**}}{(b/\Delta_M + R_M^{**} - \alpha)^2} A \\ &> 0\end{aligned}$$

□

Proof of Proposition 17. The same proof holds if I consider the fully competitive economy without bank heterogeneity, that is to say by setting $\phi^b = 0$.

– Due to adverse selection, the mass of firms undertaking H-projects increases when a productivity shock hits:

$$\begin{aligned}\frac{\partial(\bar{e}^{**} - e_{min})}{\partial \alpha} = \frac{\partial \bar{e}^{**}}{\alpha} &> 0 \\ \Delta_H \frac{p_H(p_H - \frac{r\Delta_M}{b} - \frac{r}{R_M^{**}}(1 - \frac{\Delta_M \alpha}{b})) - (\alpha p_H - c^{b=2} - r)\frac{r\Delta_M}{bR_M^{**}}}{(p_H - \frac{r\Delta_M}{b} - \frac{r}{R_M^{**}}(1 - \frac{\Delta_M \alpha}{b}))^2} &> 0 \\ p_H^2 - p_H \frac{r}{R_M^{**}} - \frac{p_H r \Delta_M}{b} + c^{b=2}\frac{r\Delta_M}{bR_M^{**}} + \frac{r^2 \Delta_M}{bR_M^{**}} &> 0 \\ p_H(p_H - \frac{r}{R_M^{**}}) + \frac{r\Delta_M}{b}(\frac{c^{b=2} + r}{R_M^{**}} - p_H) &> 0\end{aligned}$$

The first term on the left is positive since $R_M^{**} > \frac{r}{p_M}$, and the second term is positive as well if $c^{b=2} + r > p_H R_M^{**}$. But from assumption 1 $\alpha p_M > R_M^{**}$ so that the previous inequality can be replaced by $c^{b=2} + r > p_H \alpha p_M$ which holds since I have $r > \alpha p_M - c^{b=2}$, that is to say the M-yield project (no effort) cannot sustain monitoring and have a positive net present value.

Thus the mass of firms undertaking M-projects goes in the other direction and decreases when a productivity shock hits:

$$\frac{\partial(e_{max} - \bar{e}^{**})}{\partial \alpha} < 0 \tag{32}$$

– I have that, when a productivity shock hits, inefficient bank loans are less cyclical than

efficient bank loans if I have (and using the fact that I know $R_M^{**} > \frac{r}{p_M}$):

$$\begin{aligned} \frac{\partial B_M^{**}}{\partial \alpha} &< \frac{\partial B_H^{**}}{\partial \alpha} \\ \frac{R_M^{**}}{(b/\Delta_M + R_M^{**} - \alpha)^2} &< \frac{rp_H}{(p_H(\frac{e}{\Delta_H} - \alpha) + c^{pu} + r)^2} \end{aligned}$$

So a sufficient condition for the above equation to be met is for ψ to be large enough if the left hand side is decreasing in ψ, which is indeed the case, since, from assumption 3 and 4, I have:

$$\frac{\partial^2 B_M^{**}}{\partial \alpha \psi} = \frac{(b/\Delta_M + R_M^{**} - \alpha)(b/\Delta_M - R_M^{**} - \alpha)}{(b/\Delta_M + R_M^{**} - \alpha)^4} < 0 \tag{33}$$

□

Proof of Corollary 2. The heterogeneity in the fluctuations of aggregate lending across market segment or across bank type in the case of several heterogeneous banks is demonstrated as follows:

- From Proposition 17 part 1 and Proposition 16 part 2, I know that both the mass of entrepreneur choosing the efficient bank as well as individual loans offered by efficient banks increase as a result of a productivity shock. So at the aggregate, overall efficient bank lending is procyclical.
- From Proposition 17 Part 2 and Proposition 16 Part 3, I know that the mass of entrepreneur choosing the inefficient bank is countercyclical while individual loans offered by inefficient banks are less procyclical than the one offered by efficient banks. So at the aggregate, overall inefficient bank lending is less cyclical than efficient bank lending, and is procyclical or countercyclical depending on the relative force of the two opposite effects.

□

Proof of Corollary 4. From proposition 17 part 1, I obtain:

$$\frac{\partial^2 \bar{e}^{**}}{\partial \alpha \partial \phi} = \frac{r \Delta_M c^l}{b R_M^{**}} > 0$$

□

Proof of Corollary 5. From proposition 11, proposition 1 and corollary 4, I have:

$$\frac{\partial^2}{\partial\alpha\partial\phi}\left(\alpha\frac{p_H*(\bar{e}^*-e^{min})+p_M*(e^{max}-\bar{e}^*)}{e^{max}-e^{min}}\right)=\frac{p_H-p_M}{e^{max}-e^{min}}\left(\underbrace{\frac{\partial\bar{e}^{**}}{\partial\phi}}_{-}+\alpha\underbrace{\frac{\partial^2\bar{e}^{**}}{\partial\alpha\partial\phi}}_{+}\right)$$

□

Proof of Proposition 18. The several points of the proposition are demonstrated as follows:

– When a productivity shock hits, the interest rate offered to effort-making entrepreneurs is more volatile the presence of bank heterogeneity than in the social optimum case since I have:

$$\begin{aligned}\left|\frac{\partial R_H^{**}}{\partial\alpha}\right| &> \left|\frac{\partial R_H^{o}}{\partial\alpha}\right| \\ r\frac{c^{b=2}\Delta_H^2}{(p_H(\alpha\Delta_H-e)-c^{b=2}\Delta_H)^2} &> r\frac{c^l}{(p_H(\alpha-\frac{e}{\Delta_H}-\frac{c^l}{p_H}))^2} \\ \frac{(1+\phi^{b=2})}{(p_H\alpha-e\frac{p_H}{\Delta_H}-c^l(1+\phi^{b=2}))^2} &> \frac{1}{(p_H\alpha-e\frac{p_H}{\Delta_H}-c^l)^2}\end{aligned}$$

– When a productivity shock hits, the pool of effort-making entrepreneurs is less volatile in the presence of bank heterogeneity than in the social optimum case if I have:

$$\begin{aligned}\frac{\partial e^o}{\partial\alpha} &> \frac{\partial\bar{e}^{**}}{\partial\alpha} \\ \Delta_H &> \Delta_H\frac{p_H(p_H-\frac{r\Delta_M}{b}-\frac{r}{R_M^{**}}+\frac{r\Delta_M\alpha}{R_M^{**}b})-(\alpha p_H-c^{b=2}-r)\frac{r\Delta_M}{R_M^{**}b}}{(p_H-\frac{r\Delta_M}{b}-\frac{r}{R_M^{**}}+\frac{r\Delta_M\alpha}{R^{**}b})^2} \\ (p_H-\frac{r\Delta_M}{b}-\frac{r}{R_M^{**}}+\frac{r\Delta_M\alpha}{R_M^{**}b})^2 &> p_H(p_H-\frac{r\Delta_M}{b}-\frac{r}{R_M^{**}})+(c^{b=2}+r)\frac{r\Delta_M}{R_M^{**}b}\end{aligned} \tag{34}$$

Now in the extreme case where ψ becomes large, then I can ignore the terms with R_M^{**} at the denominator, so that I get, with $p_H-\frac{r\Delta_M}{b}<0$:

$$\begin{aligned}(p_H-\frac{r\Delta_M}{b})^2 &> p_H(p_H-\frac{r\Delta_M}{b}) \\ p_H-\frac{r\Delta_M}{b} &< p_H \\ -\frac{r\Delta_M}{b} &< 0\end{aligned}$$

which indeed holds. Now if ψ becomes small, the left hand side of (34) increases from assumption 4 and the right hand side decreases. So equation (34) always holds, provided that $b<\frac{r\Delta_M}{p_H}$; that is to say, if the efficient benefit of not behaving is small enough, else

if the efficient benefit is too high the inefficient bank on the M-segment of the market will have to offer relatively less attractive loans when a negative shock hits, because otherwise the entrepreneur would be better off diverting funds from a large loan when the return on behaving is low. Then as a result, with high private benefits, relatively less entrepreneurs would be attracted by M-contracts whose conditions are deteriorating faster, which would mean a larger volatility of the pool of effort-making entrepreneurs.

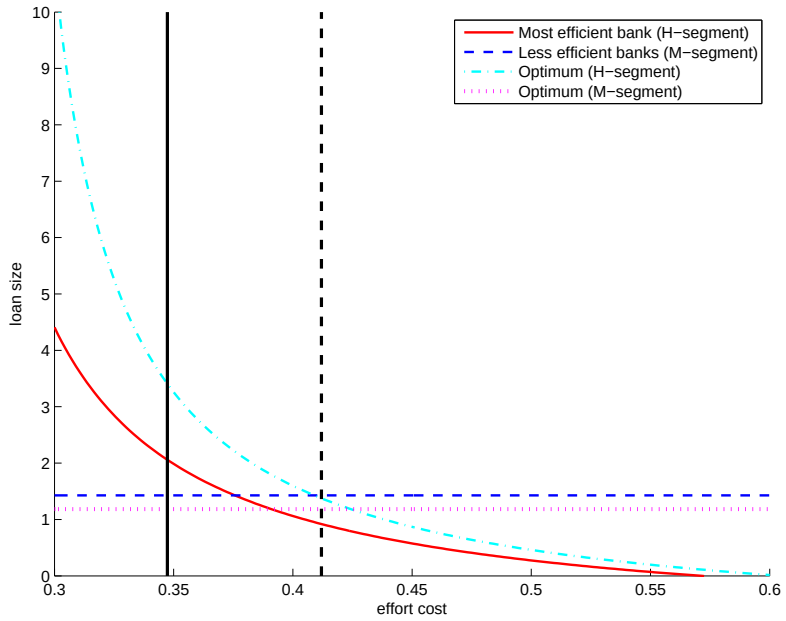
□

B Simulations

Table 1: Parametrisation

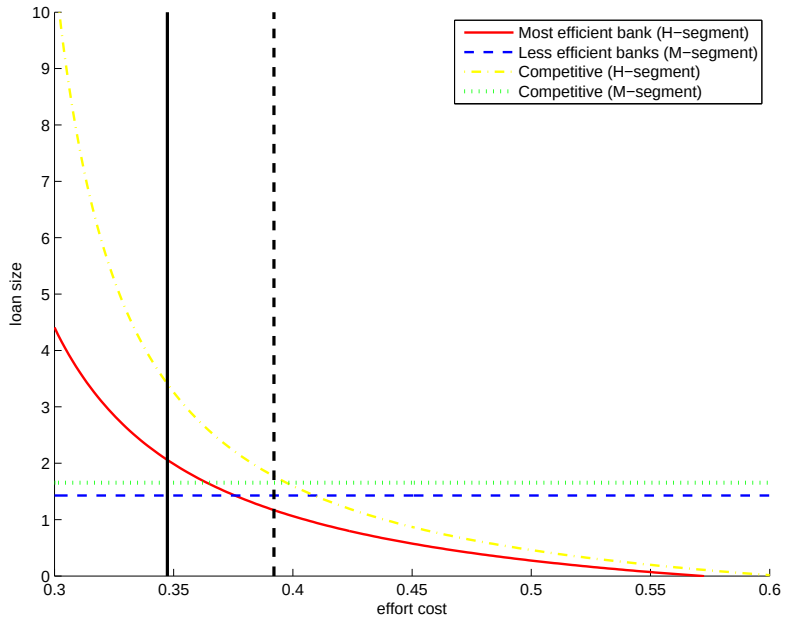
p_H	p_M	p_L	α	b	c^l	ϕ	r	A	ψ
0.9	0.6	0.4	2.24	0.24	0.2	0.5	1	1	0.1

Figure 9: Simulation of Loan size for each entrepreneur versus optimal benchmark



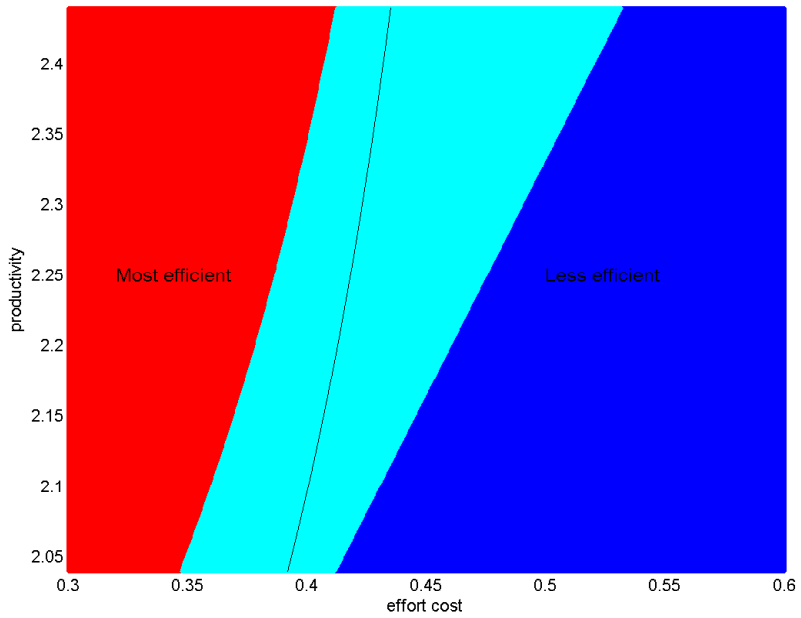
Note : for the value of α displayed in table 1. The vertical black line (resp. dashed-line) corresponds to the cut-off below which entrepreneurs self-select effort inducing contracts implementing H-projects for the heterogeneous banking equilibrium (resp. for the optimal benchmark).

Figure 10: Simulation of Loan size for each entrepreneur versus market equilibrium benchmark



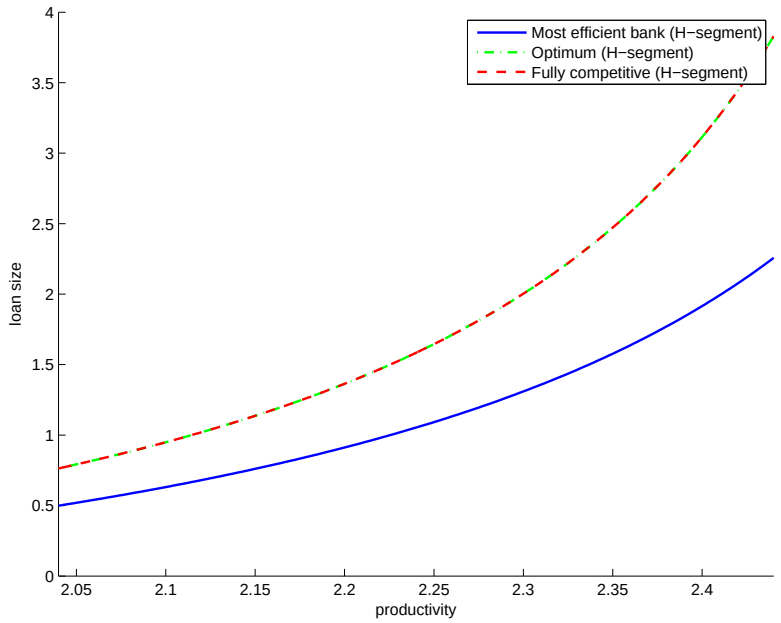
Note : for the value of α displayed in table 1. The vertical black line (resp. dashed-line) corresponds to the cut-off below which entrepreneurs self-select effort inducing contracts implementing H-projects for the heterogeneous banking equilibrium (resp. for the market equilibrium without banking heterogeneity).

Figure 11: Simulation of the State Space



Note : the red area corresponds to the efficient bank loans to effort making entrepreneurs when a inefficient bank is active; the blue and cyan area correspond to the inefficient bank loans to non-effort making entrepreneurs. The cyan area corresponds to the range of entrepreneurs which would have chosen to undertake effort at the optimum while the black line gives the cut-off entrepreneurs separating the effort and non-effort making segment of the market for the market equilibrium in the absence of heterogeneity in banking efficiency.

Figure 12: Loan size as a function of productivity : H-yield market segment



Note : loan size on the H-segment of the market is evaluated for an entrepreneur with effort cost equal to 0.3 while the cut-off effort level is 0.36.

Figure 13: Loan size as a function of productivity : M-yield market segment

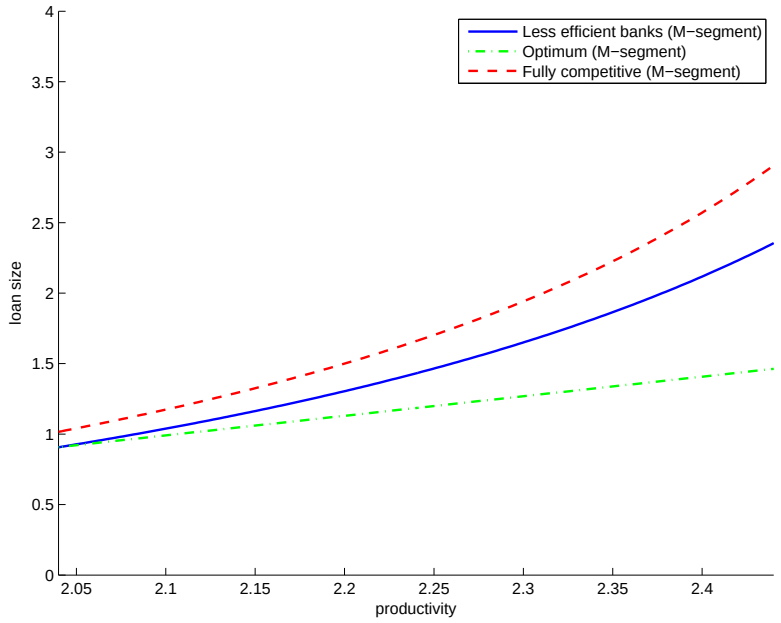


Figure 14: Average realised productivity of entrepreneurs as a function of the common productivity parameter

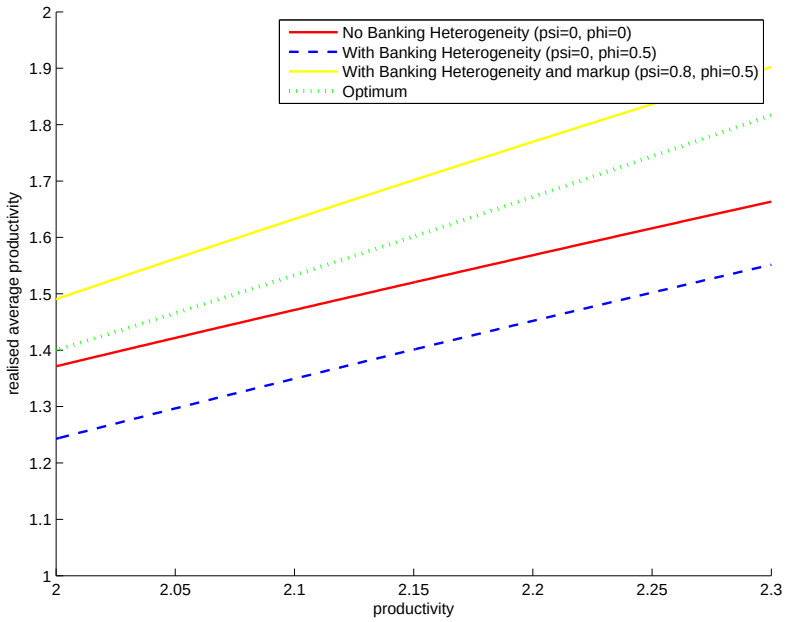


Figure 15: Weighted average realised productivity of the economy as a function of the common productivity parameter

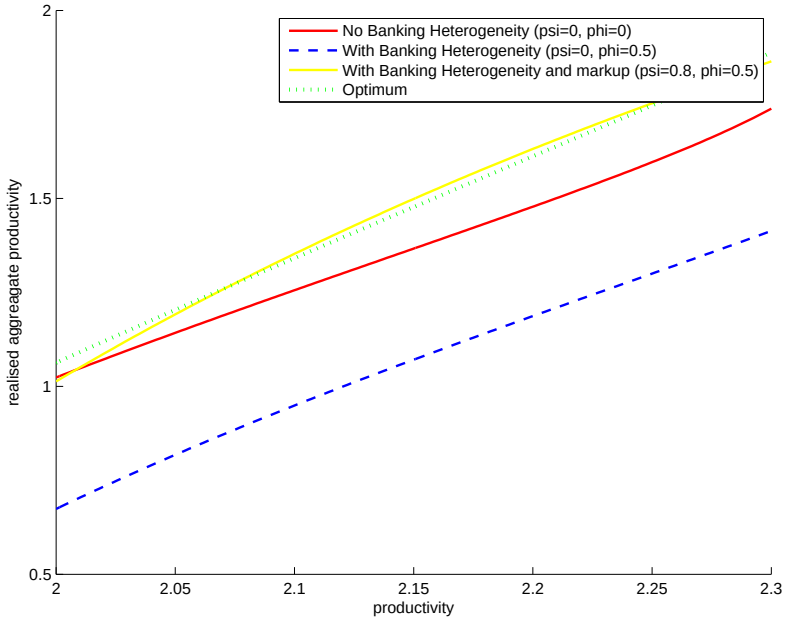


Figure 16: Aggregate lending to the economy as a function of the common productivity parameter

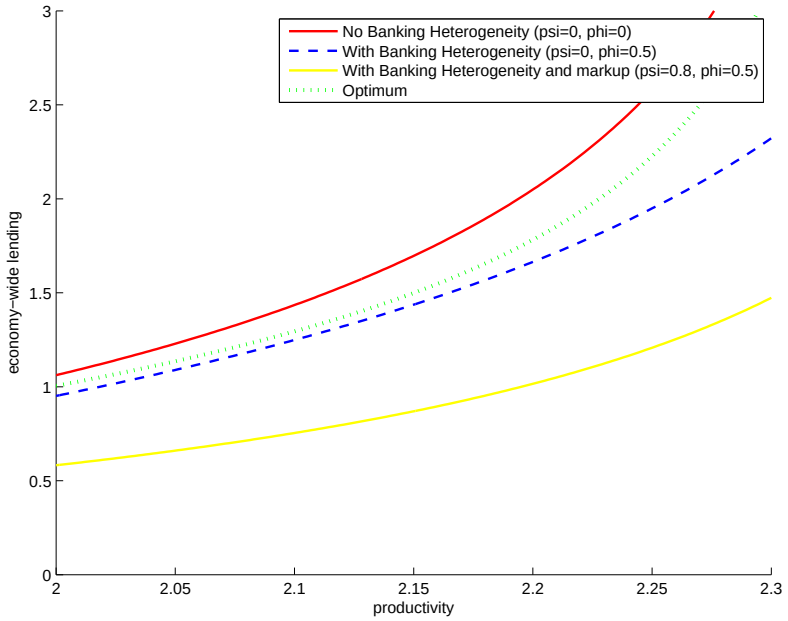


Figure 17: Aggregate output as a function of the common productivity parameter