Efficient or systemic banks: Can regulation strike a deal?

Tirupam Goel*
Bank for International Settlements

May 2024

Abstract

Should there be few large or several small banks? Large banks tend to pose scale economies, but their failure can be systemic, posing an efficiency versus financial-stability trade-off. I embed this trade-off in a macroeconomic model with heterogeneous banks, endogenous size-distribution and entry-exit, and calibrate it using micro-data. Capital regulation improves welfare by reshaping banks' size-distribution. However, regulation that equalises banks' leverage, default-rates or expected default losses is sub-optimal as it does not internalize that both efficiency and financial-stability risks are size-dependent. A hump-shaped welfare response underpins the optimal size-dependent regulation. It induces more medium-sized banks relative to the benchmark.

JEL Classification: G21, G28, L11, E44, C60

Keywords: Heterogeneous Bank Model, Size-distribution, Size-dependent Policy, G-SIBs, Financial Stability, Scale Economies.

---

*Corresponding author. ***Address:*** Bank for International Settlements, Centralbahnplatz 2, 4051, Basel, Switzerland. ***Email:*** tirupam.goel@bis.org. ***Disclaimer:*** The views expressed here are mine, and not necessarily those of the Bank for International Settlements. ***Acknowledgements:*** I am grateful to Assaf Razin, Fernando Alvarez, Pablo D'Erasmo (discussant), Karel Mertens, Eswar S. Prasad, Alejandro Rivera (discussant) and Maxim Troshkin for insightful discussions. I am also very thankful to John M. Abowd, Viral V. Acharya, Javier Bianchi, Frederic Boissay, Markus Brunnermeier, Julieta Caunedo, Dean Corbae, Christopher Huckfeldt, Benjamin Moll, Guillermo L. Ordonez, Edouard Schaal, Alp Simsek, Julia Thomas, and conference / seminar participants at Cornell, Goethe, Penn-State, CEMLA/Dallas Fed conference, Tilburg EBC conference, Cambridge MMF, Oxford IFABS, AEA, Hong Kong Economic Association conference, City University of Hong Kong, World Finance and Banking Symposium Budapest, BIS, RBI, IIM Bangalore, and Yale-NUS College for useful comments. Any errors in this paper are my responsibility. ***Disclosure:*** I declare that I have no relevant material or financial interests that relate to the research described in this paper. ***Notes:*** A previous version of the paper was titled 'Banking Industry Dynamics and Size-Dependent Capital Regulation'. Matlab codes used in this paper are available on the author's website: https://sites.google.com/site/tirupam/.

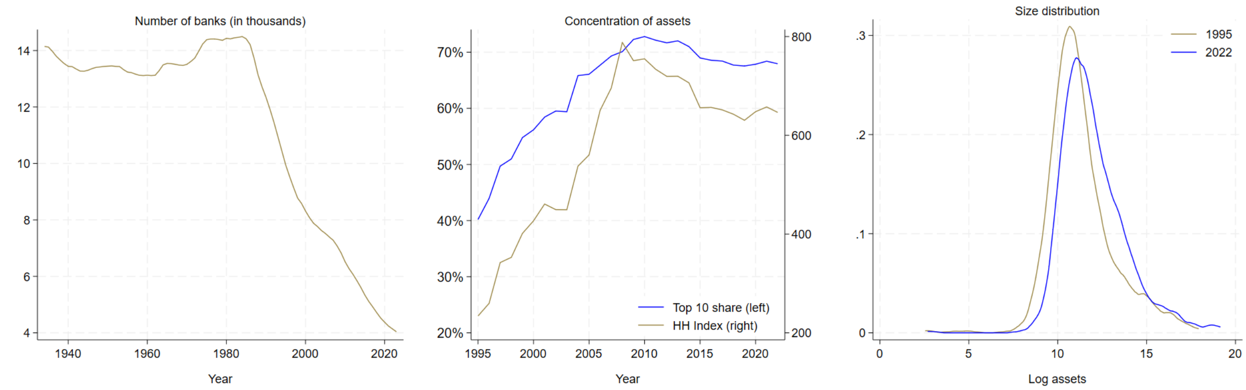

Figure 1: *First panel:* Historical count of the number of US commercial and savings banks as per the FDIC Historical Bank List database. *Second panel:* Concentration of assets among US Bank Holding Companies (BHCs) based on the Call Reports. The value of assets is deflated using the Personal Consumption Expenditure (PCE) index from FRED. HH Index stands for the Herfindahl-Hirschmann Index. A similar pattern but lower values emerge when assessing concentration at the commercial and savings banks level. *Third panel:* Kernel density of the distribution of assets deflated by the PCE index. *Sources:* FDIC Historical Bank List database, Call Reports, FRED.

1 Introduction

Should the banking sector be organized as few large or several small entities? The answer is complicated by a financial-stability versus efficiency trade-off.[1]

On the one hand, default by larger banks is more costly, raising financial stability risks. There are obvious resolution related losses associated with default like fire-sale discounts and logistical costs borne by the deposit insurer. But there are systemic losses too that are relevant in the case of default by large banks, as also evidenced during the Great Financial Crisis of 2008 [Kang et al., 2015]. Larger banks tend to be more complex and more intertwined with the financial system, which means that their default can have knock-on effects. Moreover, financial frictions can amplify the too-big-to-fail issue [Brunnermeier and Sannikov, 2014]. These concerns form the genesis of the regulatory framework for Global Systemically Important Banks (G-SIBs) [BCBS, 2018].

On the other hand, scale helps diversify risks, which can make larger banks more

[1] History illustrates the policy dilemma. The 1994 Banking and Branching Efficiency Act in the US removed hurdles and led to larger banks. But then the Great Financial Crisis of 2008 led to the introduction of too-big-to-fail reforms [BCBS, 2018] that created disincentives for banks to become larger.

efficient. Scale can also enable banks to engage in multiple business lines with synergies across them and cater to the demand for 'bundling' in financial services [Frost et al., 2019]. The rise of larger banks globally reflects these forces. In the US, for instance, there are fewer, larger, and more concentrated banks now as compared to the 1990s (see Figure 1 and also Corbae and D'Erasmo [2020]). Formally, several empirical studies, especially Hughes and Mester [2013] and Wheelock and Wilson [2018], have shown that banks exhibit increasing economies of scale – including for very large banks.[2]

In this paper, I develop a tractable framework to understand how banks should be organised given these opposing forces. Specifically, I conduct a positive and normative analysis of the role that capital regulation can play in balancing the efficiency versus financial-stability trade-off.

To provide intuition for the trade-off, I first develop a stylized model. Scale economies stem from diversification of assets, so that larger banks have a more favorable risk-return profile. However, large bank defaults are disproportionately more costly. In this case, the optimal allocation of bank capital by a benevolent planner depends on how scale economies and default losses relate to bank size. When scale economies dominate, it is better to organise the capital within a few large banks. However, when default losses are more substantial, setting up numerous small banks is better.

For the main analysis, I use the stylized model as the basis and develop a general equilibrium macroeconomic model with a heterogeneous banking sector. A crucial feature of this model is that the bank size-distribution is endogenous, and that it responds to changes in capital regulation. This creates what I call the *banking-dynamics channel* of regulation. In addition, bank defaults pose an externality that serves as a rationale for regulation. This is necessary for a welfare analysis. Finally, to ensure quantitative relevance, I calibrate the model using micro-data on US banks. In particular, I target

[2]Rents from greater market power or implicit government subsidies can also induce banks to grow larger, but these factors do not necessarily improve efficiency of financial intermediation from a social point of view and may only deliver private benefits to banks.

moments related to the distribution of banks and how bank efficiency and default losses vary with bank size.

I use the model to first conduct a positive analysis of capital regulation. A key insight is that regulation not only affects individual bank behavior, but also generates general equilibrium effects via its impact on the dynamics of the banking sector. As regulation tightens, individual banks are more constrained because capital is sticky. In response, they preserve capital by paying fewer dividends and simultaneously invest in fewer assets. In the process, banks become less leveraged and default less often. As a corollary, they spend more time in incumbency, i.e. their average age increases. That said, lower leverage also means that banks grow more slowly. The net effect of these opposing forces on the size-distribution of banks is not obvious – it depends on which force dominates quantitatively.

Then, through its impact on banking dynamics, regulation has aggregate implications. This includes an impact on overall output and welfare. In particular, because each bank retains more capital, there is an increase in aggregate capital in the steady-state. This is despite there being no possibility to raise capital externally. This underscores how the adjustment in the distribution of bank capital can somewhat mitigate the constraining effect of regulation.

To investigate the relevance of the banking-dynamics channel, I recompute the aggregate impact while keeping the bank size-distribution fixed. The decline in aggregate output in this case becomes more pronounced and tighter regulation can no longer improve welfare. This underscores that the banking-dynamics channel matters both quantitatively and qualitatively. The emphasis on the banking-dynamics channel – i.e. the heterogeneous effect that regulation can have on different banks and the aggregate implications of this – in this paper is novel relative to other studies on optimal capital regulation that rely on a representative bank model or take bank heterogeneity as given exogenously.

Building on the positive analysis, next I turn to a normative analysis of regulation to characterise the optimum. Specifically, I study the welfare implications of several counterfactual regimes that are inspired by the evolution of regulatory practice in the last two decades. First is a uniform capital-ratio requirement across banks (as in Basel I). Second is risk-sensitive requirement that ensures that banks' probabilities of default are equalised (the idea behind Basel II). A third regime equates banks' expected default losses (EL), which is a product of their probability of default (PD), loss-given-default (LDG), and exposure-at-default (EAD). This is also roughly the idea behind the G-SIB framework of Basel III.

The key insight from these analyses is that none of these regimes are actually optimal. In each of these regimes, an increase in regulatory stringency leads to an inverted U-shaped welfare response. The maximized welfare is highest in the third regime. Indeed, equating EL across banks accounts for the fact that all three aspects of banks' default risk, namely PD, EAD and LGD vary with size – which helps improve welfare. Yet, this regime falls short of internalizing the fact that both bank efficiency and financial-stability risks are size-dependent. As such, a focus on just mitigating the risks banks pose in partial. Taking into account differences in efficiency across banks is also crucial for maximising overall welfare. In other words, the optimal regime should strive to not only reduce the welfare *costs* associated with a given size-distribution of banks but also optimise the welfare *benefits*.

To this end, the paper shows that a flexible regulatory regime that varies with bank size can do better by optimising on both legs of the efficiency versus financial-stability trade-off. This regime also features a hump-shaped welfare profile, but in contrast to previous regimes, the maximized welfare is higher. In particular, this regime gives rise to more *middle-sized* banks relative to the *heavy-tailed* distribution in the benchmark economy.

Zooming in reveals three channels through which regulation impacts welfare. First, tighter regulation translates into lower financial intermediation capacity of the banking sector (ceteris paribus), which lowers output and welfare.[3] Second, regulation impacts

[3]This is despite the fact that the aggregate stock of capital increases as banks retain more earnings

the average efficiency of the banking sector by re-shaping the bank size-distribution. The sign of the impact is, however, not obvious. This is because there is a greater mass of middle-sized banks that are *more* efficient than the small-sized banks but *less* efficient as compared to the large-sized banks. Third, the impact of regulation on expected losses (EL) due to bank defaults – which ultimately affects welfare too – is also not obvious. While banks' probability of default (PD) declines as regulation tightens, change in the overall $EAD \times LGD$ of the banking sector depends on the size-distribution. Comparative statics with respect to EL underscores the welfare implications of this channel.

The net welfare effect naturally depends on the relative strengths of these channels. But the fact that these can run in opposite directions is at the core of the regulatory trade-off. It shows why focusing on some aspects (like PD or EL) of the trade-off while ignoring others (like efficiency) can be sub-optimal.

Turning to some further specifics of the model, banks are financial intermediaries that raise deposits and invest in risky assets. They operate a leveraged balance sheet that is not indeterminate (unlike in the Modigliani-Miller economy) as capital is a state-variable. Depending on idiosyncratic shocks to their assets, banks grow, shrink, or fail. New banks enter the industry with a random amount of seed capital. There is no aggregate uncertainty, and a stationary (steady-state) distribution emerges in equilibrium despite bank-level dynamics. I calibrate about half the model parameters on a standalone basis using targets from the data. The rest are estimated jointly using the method-of-moments.

A few noteworthy aspects of the calibration exercise are as follows. First is the emphasis on disciplining parameters that underpin scale economies. To this end, I target the mean and standard deviation of banks' return on assets and – importantly – how these moments vary across large and small banks. Second, I distinguish between smaller and larger banks when calibrating the loss-given-default (LGD) profile. For smaller banks, I consider LGD to correspond to the average resolution related losses incurred by a deposit insurer. For

and the equilibrium distribution of banks shifts to the 'right'.

larger banks, I refer to empirical studies (such as those based on the Great Financial Crisis) to obtain an estimate of the systemic loss associated with default. Third, I match the model-generated and empirical distributions of bank capital. Specifically, I minimise the following two metrics. One is the Kolmogorov-Smirnov (KS) statistic, which is the maximum distance between the model-generated and empirical distribution functions. Another is the absolute difference in the *power-law* exponent of the two distributions. While the KS statistic helps align the two distributions in general, matching the power-law exponents ensures that the respective heavy right-tails are aligned in particular.

The calibrated model reveals that compared to the benchmark capital regulation of 4.5%, the optimal requirement is around 5.2% for all banks.[4] The corresponding welfare gain in consumption equivalent terms is around 1%. In case of the size-dependent regime, the optimal requirement is less stringent at close to 1% in the case of small banks and stricter at around 7% for the largest banks while varying monotonically in between. This finding lends support to the G-SIB framework which also imposes stricter regulation on the larger banks. Yet, this paper calls for lower regulatory burden in the case of small banks.

A series of robustness checks confirm the qualitative takeaways. These checks also reflect that the goal of this paper is not settle the question of how strong scale economies in banking are or how systemic large banks are. Specifically, I consider variations in how efficiency and default losses are related to bank size. I use these comparative statics exercises to deduce the attendant optimal capital regulation. I find that greater default losses on average call for more stringent regulation. Greater systemic losses (which apply to larger banks) justify tighter regulation for larger banks in particular. Conversely, higher

[4] I take the Common Equity Tier-1 capital or core capital requirement in Basel III as the benchmark and derive the implications for optional regulation on that basis. This is because the G-SIB requirements – a focal point in this paper – are cast in terms of core capital. In doing so, I abstract away from any add-on capital requirements such as the counter-cyclical or capital conservation buffers. This helps explains why the optimal level in this paper is lower than other studies on capital regulation that consider the total capital ratio requirement as the benchmark.

efficiency on average leads to less stringent regulation. And when larger banks are more efficient in particular, the optimal regulatory regime going from small to large banks becomes less steep.

Finally, I consider two extensions of the model. First I endogenize asset returns. In the baseline model, asset returns do not depend on the overall size of the banking sector. In reality, when banks collectively invest in more assets, the return on any individual bank's investments is likely to be lower. Incorporating this possibility in the model shows that as banks shrink in response to tighter regulation, asset returns increase, leading to a second-round positive effect on bank behavior. This countervailing effect allows regulation to push harder. Second I endogenize the mass of banks. In the baseline model, the mass of banks is normalised to unity. However, the mass of entrants into the banking sector may vary with bank profitability. Incorporating this mechanism in the model shows that as regulation tightens, it creates disincentives for entry as bank profitability declines. As a result, regulation is more constraining relative to the baseline case and thus the optimal regulation is less stringent.

Related Literature The primary contribution of this paper is to emphasize the role of bank heterogeneity in macro-finance. Studies in this domain (e.g. Bernanke et al. [1999], Gertler and Kiyotaki [2010], Jermann and Quadrini [2012], Brunnermeier and Sannikov [2014], Boissay et al. [2016]) tend to consider a representative bank model or take bank heterogeneity as exogenously given. By contrast, this paper builds on the insight that both efficiency and default losses depend on bank size and shows that this matters for how the banking sector must be organised. In doing so, this paper embraces the idea in the seminal work by Hopenhayn [1992] who showed how heterogeneity matters for aggregate macroeconomic outcomes in the case of non-financial firms. Specifically, I develop a tractable macroeconomic model with a heterogeneous banking sector where entry-exit and size-distribution are endogenous. I use this model as a workhorse to study the design of optimal capital regulation, especially how regulation must vary by bank size.

A thin but growing literature shares this paper's focus on bank heterogeneity. An important early contribution is Corbae and D'Erasmo [2010], where the authors develop an industry dynamics model of imperfectly competitive banks to study the relation between business cycles and banking sector characteristics such as market structure, defaults, risk-taking, and loan supply. In subsequent recent work, Corbae and D'Erasmo [2021] study the effect of capital regulation on such characteristics. Focusing on the link between bank size and market power, Jamilov [2021] develops a model of monopolistic-competitive banks where large banks default less often but charge a higher markup than smaller banks, and investigates how regulation interacts with this trade-off.[5] Muller [2022] studies the joint impact of non-risk-based (i.e. leverage ratio) and risk-based capital requirements on the allocation of market shares across banks with heterogeneous productivity levels. Whited et al. [2021] documents that banks' market power varies with size and studies the heterogeneous impact of low interest rates on risk-taking. Bellifemine et al. [2022] and Wang et al. [2021] study the role of bank market power in the transmission of monetary policy, while Jamilov and Monacelli [2021] study how the distribution of market power responds to aggregate shocks. Dávila and Walther [2020] shows that implicit bailout guarantees can lead to strategic leverage spillovers from large to small banks, and Liu [2019] shows that the impact of Dodd-Frank regulation in the US is different for small versus large banks.

The present paper shares some aspects of these studies but also adds to them in the following ways. One is its emphasis on what the financial-stability versus efficiency trade-off means for the optimal organisation of the banking sector. While understandably much

[5]In Jamilov [2021], large banks are more efficient and default less often (also due to an implicit too-big-to-fail guarantee) but charge higher mark-ups. By contrast, in this paper, the expected default loss posed by larger banks is not necessarily bigger because while large banks have a lower default rate, their default is more costly. This distinction leads different policy implications. In that paper, the optimal policy is to *subsidise* all banks (especially larger banks) relative to the benchmark – while in this paper the optimal policy is to impose *stricter* regulation, especially on larger banks.

attention has been paid to competition issues, the said trade-off remains under-explored. This is despite efficiency being a significant aspect in banking, not least because of how data and technological innovations are transforming scale economies and driving financial institutions (banks and also non-banks) to become larger. Second, the paper features not only a positive but also a normative analysis of regulation. In doing so, I use global solution methods (instead of linear ones) to capture potential non-linearities in the model mechanisms. Third is the paper's calibration strategy that, as described above, takes the heterogeneity across banks seriously.

The present paper naturally also relates to a large literature on the assessment of capital regulation. An early contribution is by Van den Heuvel [2008], who shows that a mis-priced deposit insurance creates a moral hazard issue and induces banks to leverage up. This rationalizes regulation, and allows for studying the welfare effects of regulation and characterising the optimal. Relatedly, Christiano and Ikeda [2013] characterize optimal regulation when the effort that a bank exerts is unobservable by its creditors – another moral hazard issue. Malherbe [2020] studies how optimal regulation should vary across the business cycle given the trade-off that both the desire to invest and available bank capital are greater in a boom. Begenau [2020] evaluates optimal regulation in a model where households value safe assets, because of which banks' cost of funding can actually decrease when higher capital requirements make deposits scarcer. Other related papers include De Nicolo et al. [2014] that studies the effects of higher capital and liquidity requirements on bank lending, efficiency and overall welfare; Nguyen [2015] that characterizes the optimal capital requirements in the presence of government bailouts; and Zhu [2008] that studies the welfare implications of risk-weighted *viz-a-viz* non risk-weighted capital requirements.

One of the predictions in several of these papers is that optimal regulation should be tighter relative to the benchmark – this is similar to the finding in the present paper. But the underlying mechanisms are distinct. For instance, while the above studies typically use a representative-bank model or one with exogenously given heterogeneity, the present paper shows that industry dynamics, i.e. shifts in the banks' size distribution and default rates, is an important channel through which the effect of regulation transmits to the overall economy.

Another distinction relative to the literature is the emphasis on size-dependent capital requirements. After the Great-Financial-Crisis, size-dependent regulation – namely the G-SIB framework – became a core element of Basel III. The tractable heterogeneous bank model developed in this paper provides an ideal setup for the assessment of such regulation. Notably, Passmore and von Hafften [2019] complements the approach adopted in this paper. The authors use a panel generalised method-of-moments (panel-GMM) framework to show that G-SIB surcharges should be higher if social-loss-given-default is to be equalised across banks.

The paper is organised as follows. I present a stylized model to develop the intuition for the key trade-off in Section 2. Section 3 develops the main model and derives its analytical properties. Section 4 defines the stationary competitive equilibrium and discusses the inefficiency that rationalises regulation in the model. Section 5 presents the calibration strategy, the numerical solution strategy, and presents the model's quantitative properties. Section 6 focuses on the positive analysis of regulation. Section 7 focuses on a series of counterfactual policy experiments and draws implications for the optimal regulation. Both sections 6 and 7 also pursue relevant robustness checks. Section 8 extends the main model along two dimensions. Section 9 concludes. The Appendix provides data related details, explains the computational algorithms, and provides proofs of the propositions in the paper.

2 Optimal size-distribution of banks: Intuition

In this section, I present a stylized model to illustrate the social planner's dilemma when thinking about the optimal distribution of banks in an economy. Insights from this model form the basis for the main model presented in the next section.

Time is discrete and there are two dates, 0 and 1. A benevolent social planner has a fixed amount of capital K. It needs to decide the optimal number of banks to setup using this capital. Each bank can be setup on an island where it combines the allocated capital k_i with deposit funding f_i to invest in $s_i = k_i + f_i$ risky projects on date-0. Banks can only invest in projects located on their own island. Each project requires unit investment. All projects (on all islands) have the same date-1 return distribution $N(\mu, \sigma^2)$. And while projects on the same island are potentially correlated, projects from distinct islands are independent.

The distribution z_i of total return on bank i's assets is normally distributed with mean μs_i while the variance that depends on the correlation structure across assets. If the projects are perfectly correlated, the variance is given as $\sigma^2 s_i^2$, while if perfectly uncorrelated, the variance becomes $\sigma^2 s_i$. In case of negatively correlated projects, the variance could even be smaller than $\sigma^2 s_i$. A convenient method to capture the correlation structure between the projects of a bank is via a *diversification* parameter $d \in [-\infty, 2]$, and posit that

$$z_i \sim N(\mu s_i, \sigma^2 s_i^d).$$

Each bank must satisfy a minimum capital-ratio constraint: $k_i/s_i \geq \chi$. Given proportional returns on assets, the constraint is binding for all banks, and therefore, they all operate with the same capital ratio. Specifically, a bank with allocated capital k_i chooses a balance sheet of size $s_i = S(k_i) := k_i/\chi$.[6]

[6]For simplicity, I assume a non-risk weighted capital-ratio constraint that is exogenously given. In the main model, I provide a rationale for regulation and also consider variants such as risk-weighted or size-dependent regulation.

Deposit funding costs $R < \mu$ for all banks. A bank defaults if it cannot cover its deposit liabilities i.e. if $z_i < R(s_i - k_i)$. The probability of default p_i can be written as:

$$p_i = Pr\Big(z_i \leq R(s_i - k_i)\Big) = \Phi\left(\frac{R(s_i - k_i) - \mu s_i}{\sigma s_i^{d/2}}\right) \tag{1}$$

where Φ is the cumulative distribution function (CDF) of the standard Normal random variable.[7] That all banks have the same leverage implies that a bank with more capital allocation k_i would have a larger balance sheet. And because of greater diversification benefits, it would have a smaller probability of default.[8]

Finally, I assume that bank default is socially costly. I consider $\Delta(s)$ as the unit cost of default or loss-given-default (LGD), i.e. the cost incurred per unit of assets when a bank with total assets s defaults. I allow the LGD to be greater in case of larger banks: $\Delta'(s) \geq 0$, to reflect the fact that default by a larger, systemically more important bank can be disproportionately more costly.

The planner must decide the number M of banks across which to distribute the capital endowment K while maximising the net expected return of the banking sector:

$$\max_M \; NR(M) \;=\; \underbrace{\sum_{i=1}^{M}\Big(\mu s_i - R(s_i - k_i)\Big)}_{ER(M)} - EL(M) \quad s.t. \;\; \sum_{i=1}^{M} k_i = K.$$

Here $NR(M)$ is the net expected return, which is expected return $ER(M)$ net of expected losses $EL(M)$, as a function of the number of banks M in the economy. For analytic tractability, I assume that capital is distributed equally across the banks, so that

[7] As long as there are some diversification benefits, i.e. $d < 2$, the probability of default as a function of the size of the bank s_i for a given level of capital k_i goes to zero as s goes to infinity. If $d = 2$, then p_i converges asymptotically to $\Phi(\frac{R-\mu}{\sigma})$.

[8] To show this, I replace $s_i = k_i/\chi$ in Equation (1) and take the derivative w.r.t. k_i:

$$\frac{\chi^{d/2}[\frac{R-\mu}{\chi} - R]k_i^{(1-d/2)}}{\sigma} < 0.$$

$k_i = K/M = k$. And since the capital ratio constraint binds for each bank, $s_i = s = k/\chi$, i.e. the banks are also equally leveraged. As a result, $ER(M)$ is independent of M, and equals $(\mu - R)K/\chi + RK$. The planner's objective reduces to minimising $EL(M)$.

Since projects across islands are uncorrelated, bank defaults are also uncorrelated.[9] This simplifies the calculation of expected losses $EL(M)$:

$$EL(M) = \sum_{m=0}^{M} \Delta(ms)msB(m; M; p(M)) = \sum_{m=0}^{M} \Delta(ms)ms\frac{M!}{m!(M-m)!}p(M)^m(1-p(M))^{M-m}$$

Here $B(m; M; p(M))$ is the binomial probability density function which denotes the probability that m banks default at the same time with $p(M)$ being the probability of default of an individual bank when there are M banks in total.[10]

The expression for $EL(M)$ embeds an *efficiency* versus *financial-stability* trade-off for the social planner. Consider M to be relatively low so that banks are comparatively larger. Then, on the one hand, diversification benefits imply that the probability of failure of larger banks is lower, making them more efficient. On the other hand, having large banks may not be desirable socially as large bank defaults are disproportionately costly. As such, conceptually, it is not obvious as to whether larger banks should be encouraged or broken down.

[9] It is possible to generalise the model by having correlated bank defaults, such as by using the setup in Gârleanu et al. [2015]. In their paper, investors located on a circle choose projects that are also located on the circle. Projects that are closer to each other in terms of the shortest arc length between them are more correlated. Because of information frictions or acquisition costs, investors end up choosing projects that are closer to their own position on the circle. As a result, projects *within* the portfolio of an investor are more correlated than projects *across* portfolios of investors in Gârleanu et al. [2015]. As a special case of the model in Gârleanu et al. [2015] while maintaining the same spirit, in the stylized model in this paper, projects held by a bank are correlated, but projects across banks are not. I take this approach to maintain analytic tractability. Computing the expected loss function when bank defaults are correlated can be handled using the "correlated" binomial distribution, but then a closed form expression for $EL(M)$ is not available. The likely effect of correlated bank defaults on the main takeaways of the model in this section would be that the optimal number of banks would be smaller.

[10] Note that the social cost depends on the total assets of all m banks that defaulted. This is an important feature to have because when several banks default at the same time, this can reinforce default related costs, say due to fire-sale externalities. Alternatively, it is possible to assume that the social cost when m banks of size s each default is given as $m\Delta(s)s$. In this case, the computations are actually simpler, and the qualitative takeaways continue to hold.

Unfortunately, the model does not admit a general closed-form solution. To formalise the trade-off, I consider two specific functional forms for $\Delta(s)$.

First I let $\Delta(s)$ to be independent of s, i.e. large bank failures are not necessarily costlier. In this case, setting up just one bank is optimal, which makes intuitive sense as larger banks are more efficient. To see this formally, note that $EL(M)$ is given as $\Delta K p(M)/\chi$ in this case, which is minimised when $M = 1$.

Second, I consider $\Delta(s)$ to be given as δs. The expression for $EL(M)$ in this case is given as:

$$EL(M) = \sum_{m=0}^{M} \delta \left(\frac{K}{M\chi}\right)^2 m^2 B(m; M; p(M)) = \delta \left(\frac{K}{M\chi}\right)^2 \left(Mp(M)(1-p(M)) + M^2 p(M)^2\right)$$

where we use the result that if X is a binomial random variable with parameters (M,p) then $E[X^2] = Mp(1-p) + M^2p^2$. This leads to:

$$EL(M) = \underbrace{\delta \left(\frac{K}{\chi}\right)^2}_{\nu} \left(\frac{p(M) + (M-1)p(M)^2}{M}\right)$$

The profile of $EL(M)$ is not obvious. Even a general comparison of $EL(1)$ and $EL(M)$ for large M (which can then help shed light on whether one large or many smaller banks is more desirable) is not possible. Specifically, $EL(1) = \nu p(1)$, while $EL(M)$ tends to $\nu p(M)^2$ as $M \to \infty$. Depending on how flat or steep $p(M)$ as a function of M is, $p(1)$ could be higher or lower than $p(M)^2$ for large M. Moreover, setting $EL'(M)$ to zero implies that an interior solution for M may also exist.

While it is not possible to derive further insights analytically, numerical simulations help illustrate that the optimal number of banks depends on the degree of diversification. When diversification benefits are higher, the probability of failure (given the same leverage) is lower irrespective of the number of banks established (left-hand panel of Figure 2). In this case, setting up one large bank is optimal (right-hand panel). When diversification

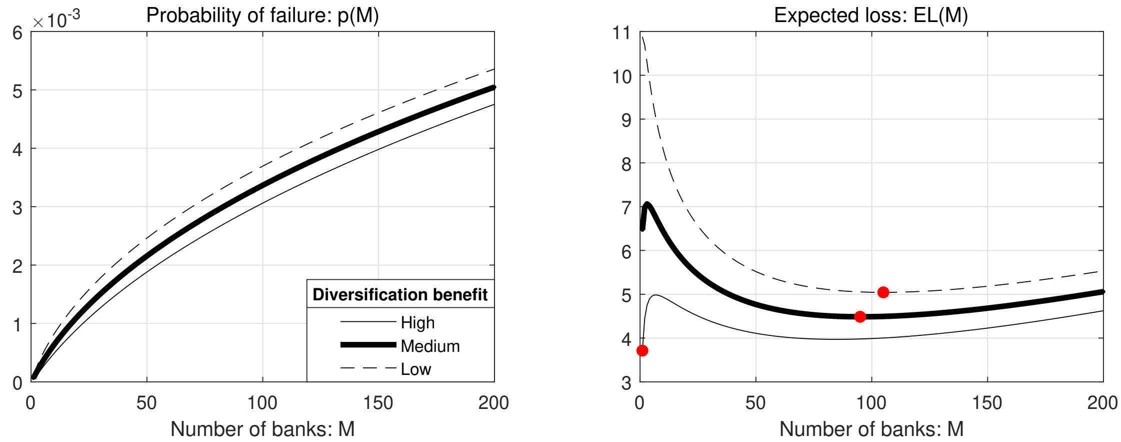

Figure 2: Probability of default (p) and expected loss (EL) as a function of the number of banks (M). The red dots in the left-hand panel show the optimal i.e. EL minimising number of banks. These illustrative computations assume that $K = 100, R = 1.04, \mu = 1.05, \sigma = 0.05, \delta = 0.1$, $\chi = 0.1$, and d taking values 1.84, 1.85 and 1.86 corresponding to high, medium, and low levels of diversification.

benefits are lower, the probability of failure is higher. Combined with the fact that large bank failure is more costly, setting up several smaller banks (instead of a few large ones) is better in this case.

The stylized model presented in this section illustrates why the choice of a socially optimal number and size distribution of banks is not obvious. It underscores that the optimal solution must take into account the fact that both efficiency and financial-stability risks posed by banks depend on their size. In other words:

Result. *The optimal number and size distribution of banks depends on how large diversification benefits are and on how much more costly large bank failures are.*

To make the exposition transparent, the stylized model uses several assumptions. For one, it uses a partial equilibrium setup where the planner dictates the allocation of capital across banks. Obviously, this is not realistic as policymakers instead use capital-ratio requirements to influence bank behavior and achieve the socially optimal outcome. In addition, while the stylized model is static, in reality banks are not one-period entities – instead they optimise over a long horizon over which they can grow, shrink, or even fail. In the main model (next section), I relax the simplifying assumptions and consider a richer framework that allows me to conduct a more realistic analysis of the trade-off highlighted using the stylized model.

3 Model

The overarching goal in this paper is to understand how capital regulation can balance the efficiency versus financial-stability trade-off by shaping the dynamics of the banking industry, and what this means for the optimal level of regulation. To this end, I develop a dynamic general equilibrium model of an economy with a heterogeneous banking sector. The model draws inspiration from two seminal papers. First is Gertler and Kiyotaki [2010], which introduces a role for banks in a standard macro-economic framework. Second is Hopenhayn [1992], which develops a model of firm-level heterogeneity. I combine elements from the macro-finance framework in the former paper with bank-level heterogeneity in the spirit of the latter paper. Naturally, in the process, I pursue innovations that help deliver an ideal model for studying the issues posed.

For one, I use elements from the stylized model in the previous section to introduce a novel efficiency versus financial-stability trade-off in the main model. Next, in contrast to Hopenhayn [1992], the balance-sheet leverage (capital structure) decision of banks (firms) in this paper is non-trivial – this allows for amplification of idiosyncratic shocks that drive exits, which is an important stylized feature in the case of banks. Moreover, a mis-priced deposit insurance and limited liability imply that banks assume higher leverage than what is socially desirable, and this rationalizes a regulatory constraint on banks' leverage. This constraint not only alters individual bank behavior, but it also affects industry dynamics and aggregate outcomes such as welfare. Taken together, these features of the model enable both positive and normative analyses of capital regulation.

Time is discrete and the horizon is infinite. The model economy consists of a household, a banking sector with heterogeneous atomistic banks, the government, and a benevolent regulator. There are bank-level dynamics, but there is no aggregate uncertainty. In what follows, I describe the various agents in the model economy.

Household The household consists of a representative worker and a unit mass of bankers. The worker receives a fixed wage income W each period – this is set to unity and serves as the numeraire. The bankers manage the banking sector and bring back dividend income E to the household. Collectively, the household consumes C and saves D in the form of bank deposits. To help keep the household' problem tractable, I assume that there is perfect consumption insurance between the two types of agents (as also in Gertler and Kiyotaki [2010]). Deposits are risk-free due to a deposit insurance scheme, and offer an interest rate R. The household is subject to a lumpsum tax T. Finally, I assume a standard constant relative risk aversion (CRRA) utility function u with a risk-aversion parameter equal to two and let β be the discount factor. The decision problem of the household, taking wage and dividend incomes as given, is as follows:

$$\max_{C_t, D_t} \quad \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(C_t) \quad s.t. \quad C_t + D_t = W_t + E_t + R_{t-1} D_{t-1} - T_t.$$

Bankers and banks Each atomistic banker manages a bank. Banks differ in terms of their *size*, that is the amount of capital n it has. They raise deposit funding d and invest in s one-period lived assets. Assets generate a risky return ψ. For tractability, I assume that the returns on assets across banks and across time are independently and identically distributed.[11] To ensure viability of banks, I focus on the case where the expected return on assets is higher than the cost of deposits: $\mathbb{E}[\psi] > R$.

Scale efficiency Following up on the intuition in Section 2, I allow larger banks to be more efficient. As discussed earlier, several empirical studies support this assumption.

[11]Considering correlated return on assets across banks or time are interesting extensions, but they increase the state-space of the bank's problem and reduce the analytical and computational tractability of the model.

For one, large banks may be able to better overcome fixed operating costs, including compliance costs [Hughes et al., 2019; Wheelock and Wilson, 2012]. In addition, they are better positioned to invest in a more diversified set of assets [Hughes and Mester, 2013; Beccalli et al., 2015]. Larger banks may also be able to offer a wider range of products with synergies between them and thus reap economies of scope [Gambacorta and van Rixtel, 2013; Baele et al., 2007; Van Lelyveld and Knot, 2009].

To capture the size-dependence of efficiency more flexibly than the approach adopted in the stylized model, I consider a specification that allows (but not necessarily imposes) larger banks to have a less volatile and higher return on assets. Specifically, I assume that bank asset returns are normally distributed and the mean and the standard deviation are functions of the bank's assets s: mean $\theta(s) = \theta_0 - \theta_1/(1+s)$ and standard deviation $\sigma(s) = \sigma_0 + \sigma_1/(1+s)$. I describe in Section 5 the calibration of these parameters using bank-level micro-data.[12]

Dividend preferences Bankers have concave preferences over the stream of dividends they pay. These preferences are not necessarily the same as the household's preferences over consumption. Two well-known regularities rationalise this distinction. First, bankers, or entrepreneurs more generally, tend to be less risk-averse as compared to workers and households (see for e.g. Kihlstrom and Laffont [1979]). Second, managers – the agents – despite having some ownership stake, may not act in the best interest of the shareholders – the principal (see for e.g. Jensen and Meckling [1976]). Similar to the approach in Bianchi and Bigio [2022], I assume that preferences over dividends e are given as $\mathcal{H}(e) = log(1+e)$. This also reflects bankers' lower risk aversion compared to the household's overall preference.

[12] The specific choice of the functional form is not crucial for the qualitative findings in this paper. The only aspect that matters is that the functions should not diverge towards infinity as s increases. Alternative functional forms such as the logistic or sigmoid functions – which are also bounded functions – lead to qualitatively findings. In particular, using these alternative forms leads to similar welfare results as described in the context of Figure 10 in Section 6.1.

Deposit insurance Banks pay an insurance premium that is proportional to the level of deposits they have. The deposit insurance fund, which is run by the government, covers the shortfall in liabilities of defaulted banks as well as any resolution related losses. In case the insurance premium is insufficient (surplus) to cover the resolution process, a lump-sum tax (subsidy) is imposed on (passed to) the household. I assume that the insurance program is mis-priced in the sense that insurance premiums do not adequately reflect banks' riskiness.[13] This, as I show below, leads to an inefficiency in banks' decisions and rationalises capital regulation.

Default and resolution A bank defaults when its capital falls below a cutoff τ.[14] Bankers have limited liability, which means they simply walk away from a failed enterprise with zero value in hand. Defaulted banks' balance sheets are resolved by the deposit insurer while ensuring depositors are paid back in full.

I assume that bank default is costly. In practice, this cost can stem from several channels. First is the obvious resolution related operational expenses incurred by the deposit insurance agency, including verification costs (see Cooley and Quadrini [2001]). Second, a forced sale of the defaulted bank's assets may fetch a discount relative to market prices, either because of intrinsic uncertainty about the quality of these assets, or because of fire-sale effects more generally (see Shleifer and Vishny [2011]). Third, default by large too-big-to-fail banks can amplify due to financial frictions [Brunnermeier and Sannikov, 2014] and lead to negative spillovers or knock-on effects [Caballero and Simsek, 2013]. This can generate losses for the wider financial system. These systemic costs may be higher during a crisis when many banks are in trouble at the same time. For the sake of brevity, I abstract from the micro-foundations of these costs. Instead, like in the stylized model, I

[13] Typical reasons for a mis-priced deposit insurance include the inability of the insurer to observe banks' risk profiles or impose risk-sensitive premium payments. See Flannery et al. [2017] for a discussion, and Van den Heuvel [2008] who uses a similar approach to justify regulation.

[14] While $\tau = 0$ corresponds to default in the strict sense, $\tau > 0$ denotes bank distress more broadly. Indeed, banks often fail before their net-worth actually falls below zero.

assume that the social loss incurred when a bank with assets s defaults is given as $\Delta(s)s$ where $\Delta(s)$ is the average unit loss per asset or loss-given-default (LGD) and s is basically the exposure-at-default (EAD). I estimate $\Delta(s)$ as an increasing function of s in Section 5.

Entry Bankers enter the industry with a random seed capital n_e. I assume n_e to be log-normally distributed as $G(\theta_G, \sigma_G)$. The log-normal distribution provides a good description of the observed size distribution of incumbent banks except for the heavy right tail (see for instance Janicki and Prescott [2006]).[15]

I assume that the seed capital is paid by the deposit insurance fund (which in turn is funded by the premiums paid by banks). While not explicitly modeled as such for brevity, this specification captures the spirit of bank entry and exit in practice where defaulted banks are rarely dissolved – instead, they are merged with a healthy bank following a bidding process. Crucially, to abstract away from adverse ex-ante incentives that this may generate (as in Dávila and Walther [2020] and Nguyen [2015] for instance), I assume that bankers do not internalise any of the post default dynamics.

Finally, it is noteworthy that in the steady state the mass of entrants and defaulters is equal and the mass of banks (and thus the mass of entrants) is fixed. That said, I consider a non-trivial bank entry problem in an extension in Section 8.2.

Capital and the regulatory constraint In line with the literature on deterrents to raising capital externally (e.g. Myers and Majluf [1984]), I assume that banks cannot issue outside equity and can only grow their capital via retained earnings. This assumption allows the capital structure of the bank to be determinate as it violates a pre-condition for the Modigliani-Miller theorem.

[15] It is useful to note here that there are generally high barriers to entry in banking (e.g. stringent process to get regulatory approval). While this could be modeled as a cost of entry (as in Hopenhayn and Prescott [1992]), in this paper that would be equivalent to deducting from the start-up capital and shifting the distribution of entrants to the left.

Limited liability combined with a mis-priced deposit insurance generate a rationale for capital regulation per the following reasoning. Deposits are risk-free and depositors do not charge a risk premium. This reduces the cost of deposit funding for the bank. Combined with limited liability, this creates incentives for the bank to assume a higher leverage than what is socially efficient. In turn, this inefficiency rationalises a minimum capital-ratio constraint on the bank.[16] A more elaborate discussion of this point is in Section 4.

I allow the regulatory constraint $\chi(s)$ to assume a general form, from being cast simply in terms of a minimum capital to asset ratio (as in Basel I), to one that depends on the riskiness of the bank (as in the case of Basel II) or the size of the bank (as in the case of the G-SIB framework, Basel III).

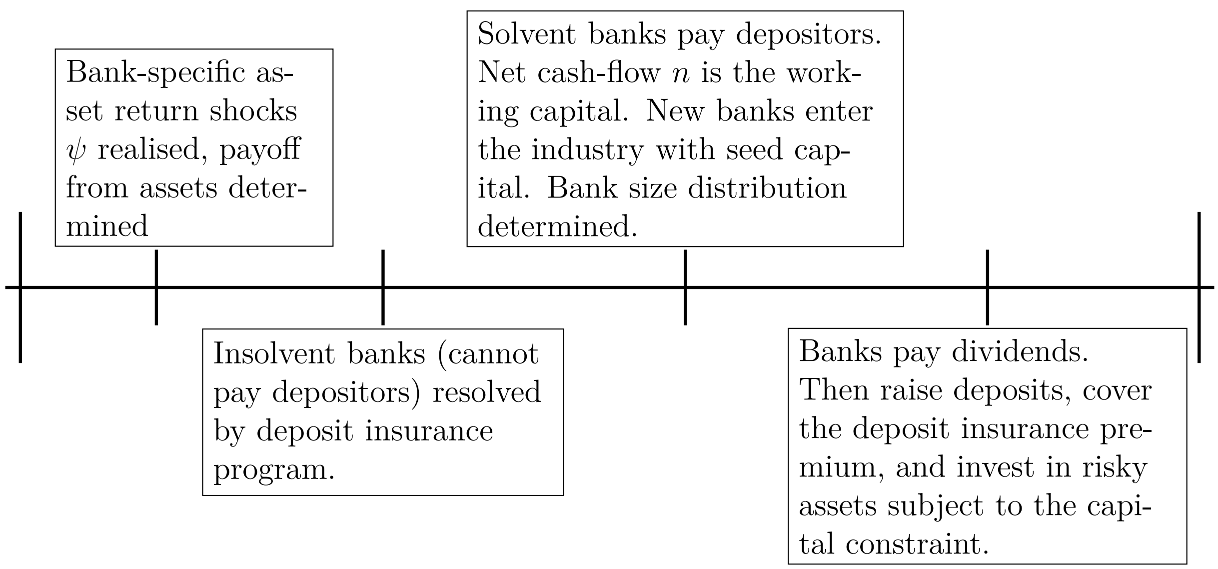

Figure 3: Intra-period sequence of events

[16]The rationale for regulation in this paper is related to that in Kareken and Wallace [1978], Santos [2001], and Van den Heuvel [2008]. A large related literature provides alternative rationales for regulation, namely fire-sale externalities [Kara and Ozsoy, 2020], and household preference for safe and liquid assets [Begenau, 2020]. Creditors (instead of regulators) may also impose a constraint due to information frictions Clementi and Hopenhayn [2006]; Goel et al. [2020], limited enforcement in Albuquerque and Hopenhayn [2004], or moral hazard Adrian and Boyarchenko [2012]; Christiano and Ikeda [2016].

3.1 Bank's problem: Recursive formulation

At the beginning of each period, bankers inherit a balance sheet, i.e. some assets and liabilities, based on the decisions taken in the previous period. The assets provide a random payoff depending on the value of the asset return shock (see timeline in Figure 3). If possible, banks cover their deposit liabilities and the remaining cash flow n becomes their operating capital. Banks that cannot cover their liabilities, or for whom $n < \tau$, exit the industry. Meanwhile, new banks enter the industry with seed capital.

At this point, banks face the following decisions: how much (i) dividends e to pay, (ii) deposits d to raise, and (iii) assets s to invest in. I setup the bank's problem recursively using capital n as the state variable.[17]

$$V(n) = \max_{s,d,e} \left(\mathcal{H}(e) + \beta \int_{\psi_c} V(n') f(\psi'; \theta(s), \sigma(s)) d\psi' \right) \quad \text{where} \quad \underbrace{\psi_c = \frac{R.d + \tau}{s}}_{\psi \text{ cutoff for defaults}}$$

$$s.t. \quad \underbrace{n - e + d = s + t.d}_{\text{Current period cash flow}} \qquad \underbrace{n' = \psi' s - R.d}_{\substack{\text{Evolution of capital} \\ \text{(Reflects balance sheet identity)}}} \qquad \underbrace{\chi \leq \frac{n-e}{s}}_{\text{Capital constraint}} \qquad 0 \leq e \quad 0 \leq d$$

Here $V(n)$ is the value function. The current period value stems from the dividends the bank pays (first term inside the bracket). Meanwhile, the expected continuation value (i.e. the integral term) depends on the bank's assets and liability decisions. Specifically, the bank combines its post-dividend capital $n - e$ with deposits d to invest in assets s and pay the deposit insurance premium $t.d$ where t is premium per unit of deposit. In turn, s and d determine the bank's capital in the next period, that is n' (where 'prime' indicates the next period). Note here that n' is a stochastic variable that equals the stochastic value of assets $\psi' s$ minus the deposit liabilities $R.d$, which is essentially reflects the balance sheet at the beginning of the next period. The value of assets depend on the asset return shock

[17] Given aggregate certainty, it is possible to assert now and verify later that the recursive formulation is well-defined. In particular, aggregates such as R and the distribution of banks need not be state variables as they will end up being constants in the steady-state equilibrium.

ψ' which has a density $f(\psi';\theta(s),\sigma(s))$ with mean $\theta(s)$ and standard deviation $\sigma(s)$. In the end, the continuation value is $V(n')$. Crucially, if n' is below τ, or equivalently the asset return shock ψ' is below ψ_c, the bank defaults, in which case the continuation value is zero.

χ is the regulatory minimum capital-ratio requirement. The constraint is expressed in terms of post-dividend capital $n-e$, which is what matters for the bank's default risk. Otherwise a bank may 'window-dress' by first reporting a higher capital ratio and then paying out a large dividend.

Properties of the bank's problem To characterise some important properties of the bank's problem, it is easier to re-write it in terms of a single decision variable e. Also, to make the analytical exposition tractable without losing generality, I assume that deposit premium is null, $t=0$, and that the default size cutoff is zero, $\tau=0$. Also note that the capital-ratio constraint is binding given proportional return on assets and the fact that expected return on assets is higher than the cost of deposits. It follows, then, that $s=(n-e)/\chi$, $d=(n-e)/(1/\chi-1)$, $\psi_c=R(1-\chi)$ and $n'=[(\psi'-R)/\chi+R](n-e)$. These together imply that:

$$V(n)=\max_e\left(\mathcal{H}(e)+\beta\int_{R(1-\chi)}V\Bigg(\Big[\underbrace{(\psi'-R)/\chi+R}_{\substack{\text{Capital}\\\text{growth}\\\text{factor (CGF)}}}\Big]\underbrace{(n-e)}_{\substack{\text{Post-}\\\text{dividend}\\\text{capital}}}\Bigg)f(\psi';\theta(s),\sigma(s))d\psi'\right).$$

The problem above captures the following trade-offs a bank faces:

1. A higher dividend payout increases current period payoff, but it reduces the current capital position, relatedly the ability to generate dividends in the future, and thus decreases the expected future value of the bank.

2. A lower capital-ratio, i.e. higher leverage, increases the bank's expected return on capital and enables it to grow faster (i.e. the capital growth factor (CGF) increases

since $\mathbb{E}\psi > R$), but it also increases the variance of its future cash flows and the probability of default.

Next, I show that the bank's problem is well defined:

Proposition 1. *There exists a unique value function $V(n)$ that solves the bank's problem, and $V(n)$ is increasing in n.*

Proof. See Appendix A. ■

Intuitively, the proof hinges on the fact that β is less than unity, that dividends cannot exceed the beginning of the period capital of the bank, and the concavity of preferences over dividends.

Given that the bank's problem cannot be solved analytically, Proposition 1 is an important *existence* result. It implies that the *Value Function Iteration* algorithm can be used to compute arbitrarily accurate estimates of V and the corresponding policy functions s, d, and e. I describe the computational method in Appendix G. Next, I present two characteristics of the bank's problem that reflect the planner's trade-off.

Proposition 2. *Banks with greater post-dividend capital $n-e$ have a smaller default rate.*

Proof. See Appendix B. ■

Intuitively, a bank with a higher $n-e$ will have more assets given the capital constraint is binding: $s = (n-e)/\chi$. The lower probability of default of such a bank follows from the fact that more assets on a bank's balance sheet implies, ceteris paribus, higher expected return on assets and a lower standard deviation. Next, consider the following remark.

Remark. *As long as $\Delta(s) > 0$, the deposit insurance agency's cost of resolving a defaulted bank – given as the total shortfall in its liabilities – is an increasing function of the size of the bank.*

Proposition 2 and the remark above together highlight the efficiency versus financial-stability trade-off that the benevolent planner faces when determining the optimal size-distribution of banks (similar as in the stylized model in Section 2): a larger bank is more efficient and has a lower probability of default, ceteris paribus, but it also poses a larger social loss conditional on default.

3.2 Distribution of bank capital

Let the cumulative distribution of bank capital be given as μ. That is, $\mu(N)$ denotes the *mass* of banks that have capital $n \leq N$ to work with in a given period. The distribution of banks evolves from one period to the other as follows (also recall Figure 3). Banks choose their balance sheet components $s(n)$ and $d(n)$ depending on their respective capital amounts n. In the next period, banks grow, shrink, or exit depending on their asset return shocks ψ. Finally, new banks enter the industry with a random seed capital distributed according to $G(.)$. Formally, the evolution is given as follows:

$$\mu'(N) = \underbrace{M' \int_{\tau}^{N} dG(n_e)}_{Entry} + \underbrace{\int_{\tau} \left(\int \mathbb{1}\Big[\tau \leq \psi' s(n) - Rd(n) \leq N\Big] f(\psi'; \theta(s(n)), \sigma(s(n))) d\psi' \right)}_{\text{Transition of incumbents net of exits}} d\mu(n)$$

The first term on the right captures the mass M' of entrant banks that enter next period and have start-up capital less then N. The second term represents the flow of incumbent banks into the $[\tau, N]$ subset of the state space, net of those that default. It is useful to express the evolution of bank capital distribution in terms of an operator: $\mu' = \mathcal{T}(\mu, M')$. $\mathcal{T}$ admits the following property:

Proposition 3. *$\mathcal{T}$ is linearly homogeneous in (μ, M). That is if μ_M is a fixed point of T corresponding to an entry mass M, $\mu_M = \mathcal{T}(\mu_M, M)$, then:*

$$\mu_M \times \frac{\hat{M}}{M} = \mathcal{T}\left(\mu_M \times \frac{\hat{M}}{M}, \hat{M}\right).$$

Proof. See Appendix C. ■

Intuitively, the key to the linear homogeneity of $\mathcal{T}$ is that the default *rate* of banks does not change with the mass of entrants. When the mass of entrants increases for a given steady state distribution, the mass of incumbents increases in a way that the mass of defaulting banks always matches the mass of entrants. Next, note the below proposition which ensures that a stationary distribution exists.

Proposition 4. *For any given $M > 0$, $\mathcal{T}$ has a fixed point.*

$$\mu_M = \mathcal{T}(\mu_M, M)$$

Proof. See Appendix D. ■

How does tighter regulation impact banks? Tighter regulation affects banks in three ways. First, a larger χ lowers the capital growth factor (CGF) of banks: $(\psi' - R)/\chi + R \downarrow$. Second, it reduces the default cutoff and thus lowers the default rate of banks: $\psi_c = R(1 - \chi) \downarrow \implies Pr(\psi' \leq \psi_c) \downarrow$. Third, banks lower their dividend payouts. This is because as the capital constraint binds more strongly, the opportunity cost of paying dividends (i.e. distributing capital) increases.

The combined effect of these mechanisms on the dynamics of the banking industry – the size distribution in particular – is not obvious *ex-ante*. To see this, consider the expected capital growth factor (ECGF) of a bank of size n:

$$ECGF(n) = \int_{R(1-\chi)} \left(\frac{\psi' - R}{\chi} + R \right) f(\psi'; \theta(s(n), \sigma(s(n))) d\psi'$$

The ECGF can increase or decrease as χ increases. This is because as regulation tightens the growth factor becomes smaller but the probability of survival increases. Nonetheless, assuming for the sake of simplicity that the distribution of ψ does not depend on s leads to the following result.

Proposition 5. *If the distribution of ψ does not depend on s, then tighter regulation lowers the expected capital growth factor (ECGF) of banks.*

Proof. See Appendix E. ■

Still, the impact of tighter regulation on how the net worth of banks evolves conditional on survival is not obvious. For a bank with current period capital n, the capital position in the next period is given as:

$$n' = [(\psi' - R)/\chi + R]\,(n - e).$$

As regulation tightens, the capital growth factor (CGF) becomes smaller (as discussed above), but banks pay less dividends e so that $n - e$ increases. This makes the impact ambiguous. Part of the problem in signing the impact is that a closed form solution for banks' dividend policy function $e(n)$ is not available. That said, intuitively, for larger banks (as compared to smaller ones), cutting dividends is less costly due to the concavity of preferences over dividends. Thus, larger banks are likely to experience a smaller impact on their growth prospects.

It is typical in heterogeneous agent models that closed form solutions are not available. This limits the ability to obtain further analytical insights. As a result, I use numerical methods to solve the model in Section 5 and conduct counterfactual policy experiments on that basis. Appendix G provides further details of the computational methods.

4 Stationary competitive equilibrium

I focus on the stationary competitive equilibrium (SCE) of the economy where despite bank level dynamics, aggregates – including the size-distribution of banks – are time-invariant.

Definition For a given capital constraint χ, an SCE consists of (i) bank value function $V(n)$, (ii) bank policy functions $s(n), d(n), e(n)$, (iii) bank capital distribution $\mu(n)$, (iv) entrant mass M, (v) aggregate bank capital N, bank dividends E, consumption C, deposits D, output Y, taxes T, default costs O and interest rate R such that:

1. $V(n), s(n), d(n)$ and $e(n)$ solve the bank's problem given R;

2. C satisfies the household's first-order-condition given R;

3. Deposit market clears at interest rate R:

$$\int d(n) d\mu(n) = D; \tag{2}$$

4. Goods market clears: $W + Y = C + S + O$ where:

$$\begin{aligned} \text{Output:} \quad & Y = E[\psi]S \\ \text{Consumption:} \quad & C = E + W + (R-1)D - T; \\ \text{Dividends:} \quad & E = \textstyle\int e(n) d\mu(n); \\ \text{Bank Assets:} \quad & S = \textstyle\int s(n) d\mu(n); \\ \text{Default cost:} \quad & O = \textstyle\int \left(\int^{\psi_c} \Delta(s(n)) \psi' s(n) f(\psi'; \theta(s(n)), \sigma(s(n))) d\psi' \right) d\mu(n); \end{aligned}$$

5. The distribution of bank capital is the unique fixed point of the distribution evolution operator $\mathcal{T}$ given entrant mass M:

$$\mu = \mathcal{T}(\mu, M);$$

6. And the government runs a balanced budget:

$$T + tD = M \int n_e dG(n_e) +$$

$$\int \left(\int^{\psi_c} \Big((1 - \Delta(s(n)))\psi' s(n) - Rd(n) \Big) f(\psi'; \theta(s(n)), \sigma(s(n))) d\psi' \right) d\mu(n)$$

where the left-hand side terms denote the lump-sum tax and deposit insurance premium proceeds respectively, while the right-hand side terms denote the total start-up funding cost and the shortfall in liabilities of defaulted banks respectively.

The existence of an equilibrium is facilitated by the fact that the bank's problem is well defined, admits unique value and policy functions, and that an invariant distribution of bank capital exists. The equilibrium can be solved for as follows. First, the household's first-order condition implies that $R = 1/\beta$ since C is time-invariant in a stationary equilibrium. Given R, the bank's problem is solvable, and policy functions are determined. In turn, the steady-state distribution of capital as well as the equilibrium mass of entry-exit are obtained. Finally all other aggregates are pinned down using the expressions noted in the definition of the SCE above. I note this existence result in the proposition below:

Proposition 6. *Given a capital constraint χ, the model economy admits a unique stationary competitive equilibrium.*

Rationale for regulation Before closing this section, a few comments on the social efficiency of the stationary competitive equilibrium (SCE) are in order. The goal is to show that there is a role for regulation in the model, that regulation has meaningful welfare implications, and that a normative analysis of regulation is possible.

To this end, I compare the problem of a constrained social planner with that of the banks, and provide intuition for why the planner's choices differ from the banks' privately optimal choices. I consider a planner that wishes to maximize the lifetime utility of the representative household in the steady-state: $u(C)/(1-\beta)$, but is constrained in its planning abilities in the following sense. While it can dictate decision rules $s(.), e(.)$ to incumbent banks, it does not interfere with government's budget constraint or dictate consumption rules to the household. Planner's decision rules $s(.), e(.)$ map to household consumption C exactly like in the SCE, that is: $C = W + Y - S - O = W + (E[\psi]-1)S - O$. Since W is given exogenously, the planner's problem is effectively:

$$\max_{s(n),e(n)} (E[\psi]-1)\int s(n)d\mu(n) - \int \Big(\int^{\psi_c} \Delta(s(n))\psi' s(n) f(\psi';\theta(s(n)),\sigma(s(n)))d\psi'\Big)d\mu(n)$$

Note that the choice of $s(n), e(n)$ by the planner shapes $\mu(n)$, and ultimately affects both the value of bank intermediation (first term) as well as the default losses (second term). The above expression thus highlights the wedge between the planner's objective (which takes into account default losses) and the banks' objectives (which ignores default losses). I refer to this wedge as the *bank default externality.* This externality stems from the fact that banks assume higher leverage (than what is socially optimal) due to a mispriced deposit insurance, which in turn increases their default probabilities, and therefore increases the social loss associated with their default which they do not internalize. The default externality rationalises regulatory intervention in this model.[18]

5 Quantitative analysis of the stationary equilibrium

In this section, I first describe the calibration of the benchmark model. I then solve for the stationary competitive equilibrium of the model economy numerically.

5.1 Calibration

To calibrate the model, I create a panel dataset on US savings and commercial banks for the period 2000 to 2022. The dataset contains various balance sheet and income statement items at an annual frequency. See Appendix F for further details regarding the

[18]Proving formally that the stationary equilibrium is constrained inefficient is beyond the scope of this paper. The main reason is that closed form solutions for bank policy functions and equilibrium distribution of capital are not available. In the next section, I confirm the intuition provided here by showing via numerical methods that tighter bank regulation can indeed lead to non-trivial welfare improvements in this economy.

data sources, data processing steps, variable definitions, and summary statistics.

Parameters (value set individually)	Symbol	Value
Discount factor	β	0.99
Deposit insurance premium	t	20 bps
Benchmark regulation	χ	4.5%
Loss-rate for failed banks: resolution costs (percent of assets)	ι	22%
Loss-rate for failed banks: systemic costs (percent of GDP)	δ	[23%,63%]
Parameters (value set jointly using Method of Moments)	**Symbol**	**Value**
Mean asset pay-off: common component	θ_0	1.0201
Mean asset pay-off: size dependence	θ_1	.0051
Stdev of asset pay-off: common component	σ_0	.0195
Stdev of asset pay-off: size dependence	σ_1	.0055
Mean size of entrant	θ_G	165.02
Stdev of size of entrant	σ_G	7.4954
Default threshold	τ	7.0114
Moments	**Data**	**Model**
Mean of ROA	0.760%	0.803%
Stdev of ROA	0.724%	2.208%
Mean of ROA, larger versus smaller banks	17.3 bps	27.5 bps
Stdev of ROA, larger versus smaller banks	-32.7 bps	-29.7 bps
Dividend payout to capital ratio	4.61%	3.60%
Exit rate	3.96%	2.46%
Ratio of smallest to median bank	1.45%	1.03%
KS statistic (relative to data distribution)	0.0	0.0515
Power-law exponent of bank-size distribution	-0.7646	-0.7288

Table 1: Summary of parameter values (first two blocks), and a comparison of data and model moments (third block). ROA is the return on assets, i.e. the ratio of net income during an year to the end of year assets. Stdev is the standard deviation, which is computed for each bank over the full sample period period. Dividend payout ratio is the ratio of divided paid during an year to the end of year capital. ROA, its standard deviation, and the dividend payment ratio are winsorized at the 1st and 99th percentiles to control for the effect of extreme outliers. Banks are divided into small and large based on the median size cutoff. Exit rate is the ratio of annual count of defaults (including mergers) to the number of incumbent banks, averaged over the sample period. KS stands for the Kolmogorov-Smirnov statistic, which is equal to the maximum distance between the model and data implied cumulative distributions of bank capital. The maximum distance is computed using the point-wise distance between the two distributions on a grid. See appendix G for details on the grid. The total sample size is 149,280 observations on 10,604 banks. See Appendix F for more details on the data procurement and processing.

For the calibration, I divide the model parameters into two sets. The first set of parameters are calibrated on a standalone basis (first block in Table 1) . I set the discount factor β to 0.99, which is consistent with an annual risk free interest rate of around 1% and is common in the literature. The deposit insurance premium rate is set to 20 basis points (bps), which is guided by the FDIC deposit insurance premium rate that typically varies between 2 to 40 bps.

I set χ to be 4.5%, in line with the Basel III minimum common-equity Tier-1 capital-ratio requirement. It is useful to note here that some other studies (e.g. Begenau [2020], Corbae and D'Erasmo [2021]) have used a higher minimum requirement close to 8% in order to account for Basel III's add-on requirements such as the counter-cyclical and capital conservation buffers. By comparison, this paper takes the *core* Basel III requirement as the benchmark and derives predictions for what the optimal requirement relative to that benchmark. The G-SIB requirements are also cast in terms of the core capital ratio, which facilitates a comparison of the predictions in this paper and the G-SIB surcharges.

Next, to calibrate the unit loss-given-default function $\Delta(s)$, I distinguish between small and large banks. The average resolution related loss incurred by the FDIC is around 22% of the defaulted bank's assets. For small banks, this is a good proxy of the social cost of failure. For larger banks, however, the repercussions of a default can be *systemic*, say due to knock-on effects. Measuring these costs is not easy. A meta study by the Basel Committee on Banking Supervision (BCBS) reports that the estimated present-discounted cost of banking crises globally ranges widely between 16% to around 300% of GDP, with a median of 63% BCBS [2010]. The IMF banking crises database [Laeven and Valencia, 2013] also finds a wide range of estimates but a lower median cost of 23% of GDP. For the US, Atkinson et al. [2013] estimates the cost of the 2008 crisis to be between 40-90% of GDP. These estimates are noisy, but they underscore that the social losses in case of large bank failures can be disproportionately larger.

Motivated by these observations, I consider the following functional form for $\Delta(s)$:

$$\Delta(s) = \iota + \left(\frac{\delta Y}{s_{max} - s_{threshold}}\right)\left(\frac{s - s_{threshold}}{s_{max} - s_{threshold}}\right)\mathbb{1}(s > s_{threshold})$$

This form captures the idea that the resolution related loss ι applies to all banks and

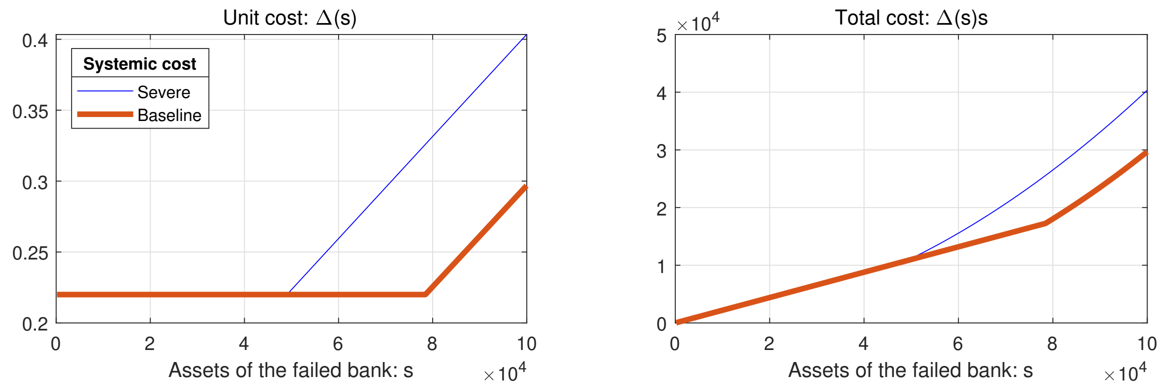

Figure 4: The first panel shows two alternative profiles for the unit cost of bank default $\Delta(s)$, while the second panel show the total cost $\Delta(s)s$. The baseline scenario assumes $s_{threshold} = 0.8s_{max}$ and the systemic cost component to be $\delta = 23\%$. The severe scenario assumes $s_{threshold} = 0.5\ s_{max}$ and the systemic cost component to be $\delta = 63\%$. In both scenarios, the cost of resolution is assumed to be the same, i.e. 22% of assets.

equals 22% of their assets while the systemic loss δ only applies to banks proportionally above a certain size threshold $s_{threshold}$. Specifically, the systemic loss is zero for banks at the threshold and increases with bank size thereafter. In line with the headline estimates in the IMF and BCBS studies, I let the systemic loss δ to range between 23% and 63% of the total economic output Y in case of default by the largest bank of size s_{max}. Two potential profiles for $\Delta(s)$ are illustrated in Figure 4.[19]

The second set of parameters are estimated jointly using the Method-of-Moments (second block in Table 1). These parameters determine the profile of banks' return on assets, namely $\theta_0, \theta_1, \sigma_0, \sigma_1$, the distribution of start-up capital, namely θ_G, σ_G, and the default threshold τ. To estimate these, I target a number of moments computed from bank-level micro data. The goal is to use moments that are informative about how efficiency varies by bank size and how the bank size distribution is shaped.

The first two moments are the mean and standard deviation of return on assets (ROA). These relates to the overall profitability of banks. Both these moments are based on pooled

[19]In principle, Y and s_{max} are endogenous, which also makes $\Delta(s)$ endogenous. In this case, an iterative algorithm is necessary to solve for the equilibrium (i.e. guess Y and s_{max}, solve, adjust the guess, re-solve, and so on). However, to avoid this complication, I fix Y and s_{max} at their respective values in an economy without systemic cost of failure i.e. $\delta = 0$.

bank-level data, wherein one period in the model is considered equivalent to one year in the data. Next, to discipline the profitability of larger banks *relative* to small banks in the model, I target the difference in mean ROA between larger and smaller banks (which is positive) and the difference in standard deviation of ROA (which is negative). For this, I classify banks as larger or smaller based on the median bank size. Figure 19 in Appendix F provides an illustration of how ROA and its standard deviation varies with bank size. Next, I target the dividend payout to capital ratio, which is another gauge of bank efficiency. I then target the exit rate which is a key aspect of the dynamics of the banking industry. I define exit rate as the ratio of annual number of defaults (including mergers) to the number of incumbent banks, averaged across the sample period.

Finally, I ensure that the model-implied and empirical bank-size distribution are well aligned. For this, I adopt a three-pronged strategy. A first moment that I target in this regard is the ratio of the size of the smallest and the median banks. Second, I minimise the Kolmogorov-Smirnov (KS) statistic that captures the maximum distance between the empirical cumulative distribution function obtained from the model-implied distribution and from the data. Third, I minimize the distance between the power law (PL) exponent estimated on the model-implied distribution and its empirical counterpart.[20] Together, these three moments help ensure that the model-implied distribution is as close as possible to the empirical one not just in terms of a few select moments but also in terms of its overall shape, especially its heavy-tail which is an important empirical regularity in banking. I pursue these minimizations as part of the Method-of-Moments estimation.

The third block in Table 1 describes the results of the Method-of-Moments estimation. Overall, I estimate 7 parameters using 9 moments. It is useful to note that two of the targets, namely the KS statistic and the PL exponent, effectively reflect multiple moments of the bank size distribution as opposed to some specific ones like the median or the

[20] To estimate the PL component, I regress the log density on the log size of banks for when size is above a certain threshold. The threshold itself is chosen so that the Kolmogorov-Smirnov goodness-of-fit statistic is optimised.

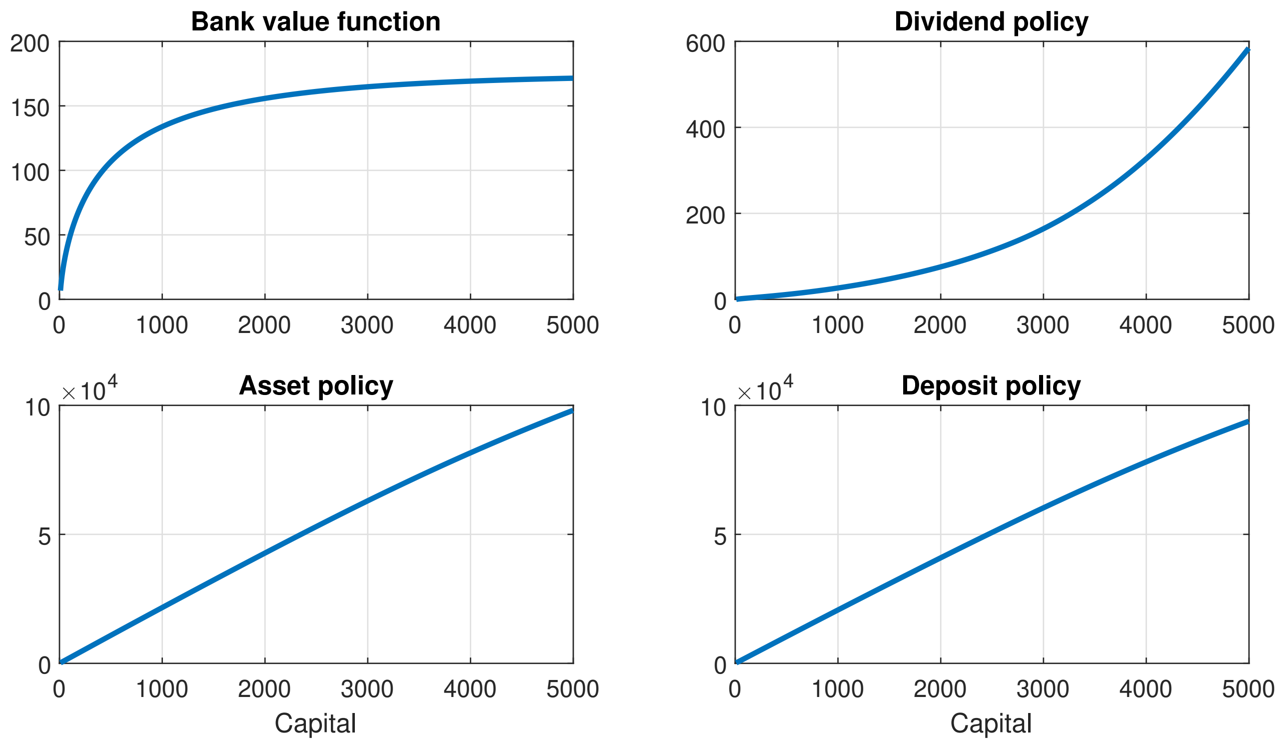

Figure 5: Bank value and policies as a function of capital, which is the state variable.

mean. As such, it is natural that the data and model moments are not exactly equal. Nevertheless, the estimation ensures that they are reasonably consistent and that the model is well disciplined by the data.

5.2 Assessing the stationary competitive equilibrium

I solve the stationary competitive equilibrium of the model economy using global solution methods – these are detailed in Appendix G.[21] The equilibrium value and policy functions of the banks are plotted in Figure 5. The value function V is concave and increasing in bank capital n, which is the state variable (panel 1). This is expected, not least given the concavity of preferences bankers have over dividends. Dividend policy is shown in panel 2. The convexity of dividend policy underscores that when capital is low, banks choose a lower dividend to capital ratio given that capital is more valuable to preserve (in order to

[21] Matlab codes used to solve the model are available on the author's website.

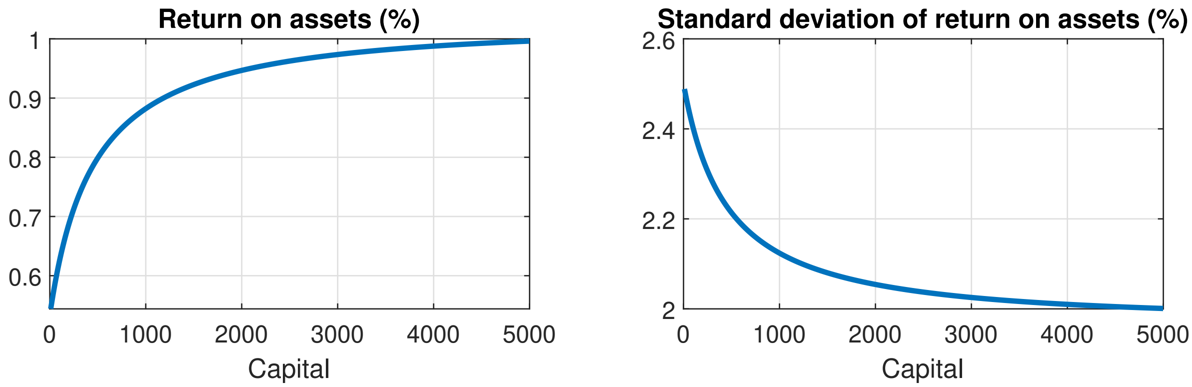

Figure 6: Variation in bank profitability across banks of different sizes.

meet the capital constraint). A larger bank, by contrast, can afford to pay an increasingly larger fraction of their capital as dividends because the marginal value of retained earnings is lower (due to the concavity of the value function). Panels 3 and 4 show that because banks are capital constrained by regulation, those with more capital are able to acquire more deposit funding and more assets.

A key feature of the model is to incorporate differences in bank efficiency by size. Figure 6 shows how mean return on assets (ROA) and its standard deviation vary with bank size: larger banks have a higher and less volatile ROA. This not only reflects the empirical regularity that underpins our calibration, but is also consistent with empirical studies on scale economies in banking, such as Wheelock and Wilson [2018], and Hughes and Mester [2013] who show that even after adjusting for potentially greater risk-taking, bigger banks pose higher efficiency.

Another key aspect of the model is the endogenous size distribution of banks. In Figure 7, I compare the model generated distribution with that in data. The first panel shows the cumulative distribution function. The close alignment of the model and empirical distribution is not unexpected given that the distribution is one of the targets of the calibration. But this is not guaranteed *ex-ante* either, and reflects the strength of an otherwise parsimonious model.

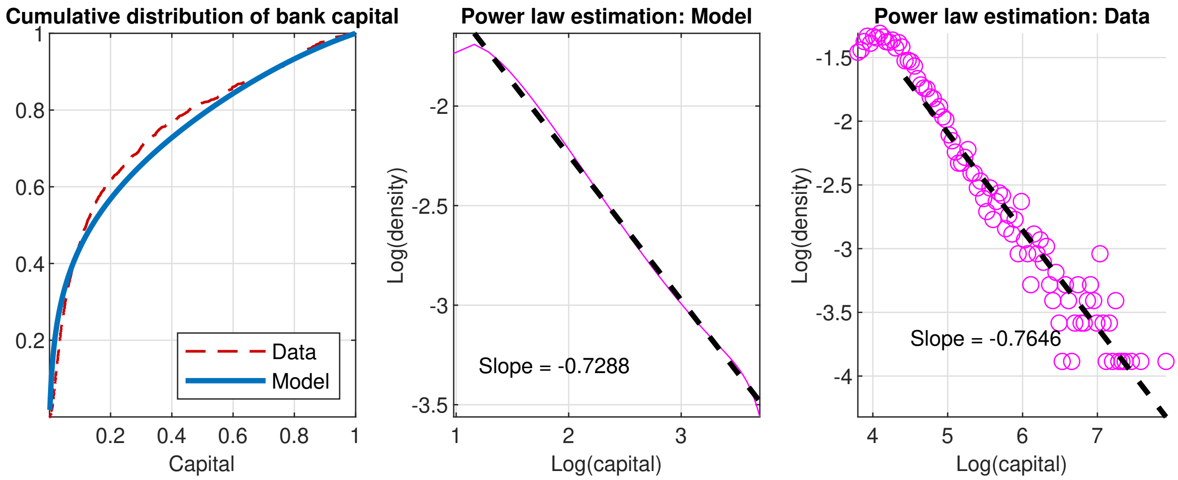

Figure 7: A comparison of model-implied distribution of bank capital with its empirical counterpart i.e. the pooled empirical distribution of banks during 2000-2022. The first panel compares the cumulative distributions, where the x-axis is normalised to the range from 0 to 1. The second and third panels estimate the power law exponent for the model and data distributions respectively.

The next two panels of Figure 7 estimate a power law on the right-tail of the model and data distributions respectively. These panels underscore the ability of the model to not only generate a heavy tailed distribution of bank capital, but also one that is closely aligned with its empirical counterpart (compare the slopes of the dotted fitted line in the respective panels).

A good match of the model-implied and empirical distributions serves two important purposes. It allows for a quantitatively relevant assessment of the effect of changes in regulation on banking industry dynamics. Relatedly it facilitates an understanding of what changes in banking dynamics implies for optimal regulation. I pursue these analyses in the next section.

6 Banking sector response to tighter regulation

In the benchmark economy, all banks face the same non-risk-sensitive minimum capital-ratio requirement χ of 4.5%. The goal in this section is to understand how a change in

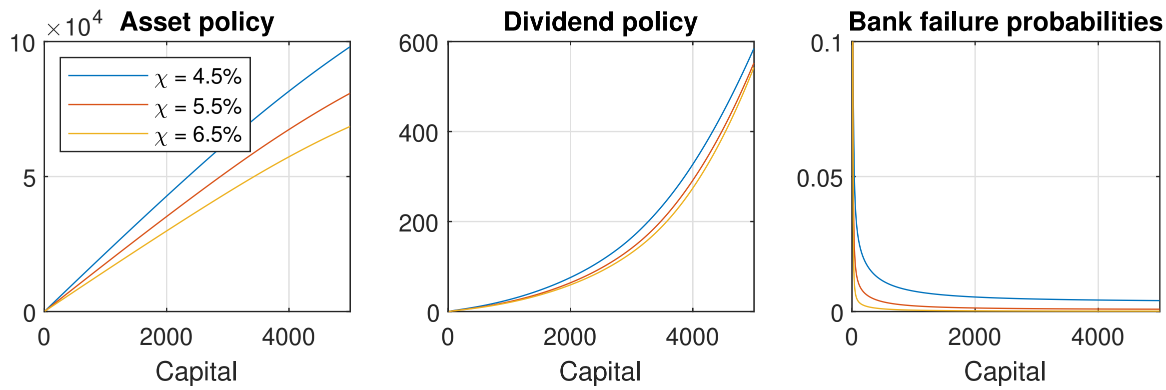

Figure 8: Response in banks' policy functions and default probabilities as a function of the minimum capital-ratio requirement χ.

the capital requirement affects the organisation of the banking sector, and to derive the optimal regulation.

As χ increases uniformly for all banks, they are naturally more constrained and are able to invest in fewer assets (first panel of Figure 8). At the same time, banks pays smaller dividends (second panel). This is because as regulation tightens, the shadow value of capital increases, and retained earnings become more important. In addition, banks' default probabilities decline (third panel). Indeed, as banks become less leveraged, deposit liabilities relative to expected payoff from assets become smaller, and so does the likelihood that banks are unable to cover their liabilities.

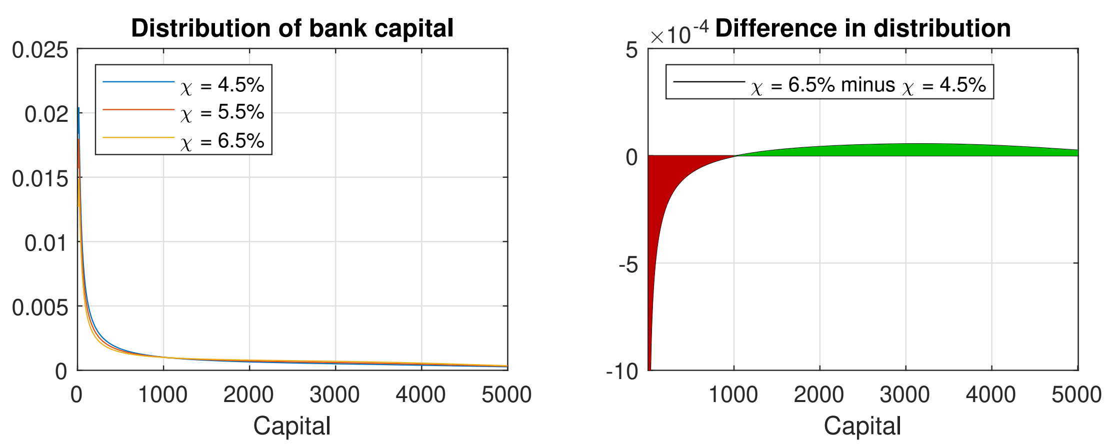

Figure 9: Response in bank distribution as a function of the minimum capital-ratio requirement χ.

The impact of tighter regulation on the distribution of banks is more interesting (first panel of Figure 9, also recall discussion in Section 3.2). On the one hand, since banks default less often, they spend more time in incumbency and their average age increases (Proposition 2). On the other hand, since banks are less leveraged, they grow at a slower rate (Proposition 5). The combined effect is that as regulation tightens, the mass of middle-sized banks increases (second panel).

6.1 Optimal regulation

The welfare implication of tighter regulation is not obvious. There are three channels that work in potentially different directions and pose trade-offs. First, tighter regulation reduces banks' probability of default (PD), which is welfare improving. Second, tighter regulation makes each bank more constrained, and leads to lower bank intermediation per unit of capital, which is also welfare reducing. Third, as a result of the rightward shift (in the first order stochastic dominance sense) in the distribution of bank capital, average efficiency of the banking sector increases (welfare improving) but at the same time the aggregate exposure at default (EAD) increases (welfare reducing).

To assess which effect dominates, I consider the problem of a benevolent regulator that strives to maximize overall welfare by adjusting minimum capital-ratio requirement χ. Welfare in this economy is measured by the household's lifetime utility $u(C)/(1-\beta)$ where C is the aggregate consumption of bankers and workers in the household. I focus on comparing welfare across steady states, i.e., before and after an unanticipated change in regulation.[22]

[22] I abstract away from welfare dynamics during the transition from one steady-state economy. In large part, this is due to the computational challenges associated with computing the transition of the entire distribution of bank capital. For example, it is not obvious if the approach in Krusell and Smith [1998] (where the entire distribution can be adequately summarized by a few moments) can be used given that the bank distribution in this paper is heavy tailed. A potential caveat of comparing welfare in steady-states is that banks' adjustment to new regulation may initially lead to lower welfare, which can make the transition to the new steady-state prohibitively costly. However, given that in practice major reforms are typically phased-in gradually, giving banks time to adjust, this caveat is likely to be less relevant.

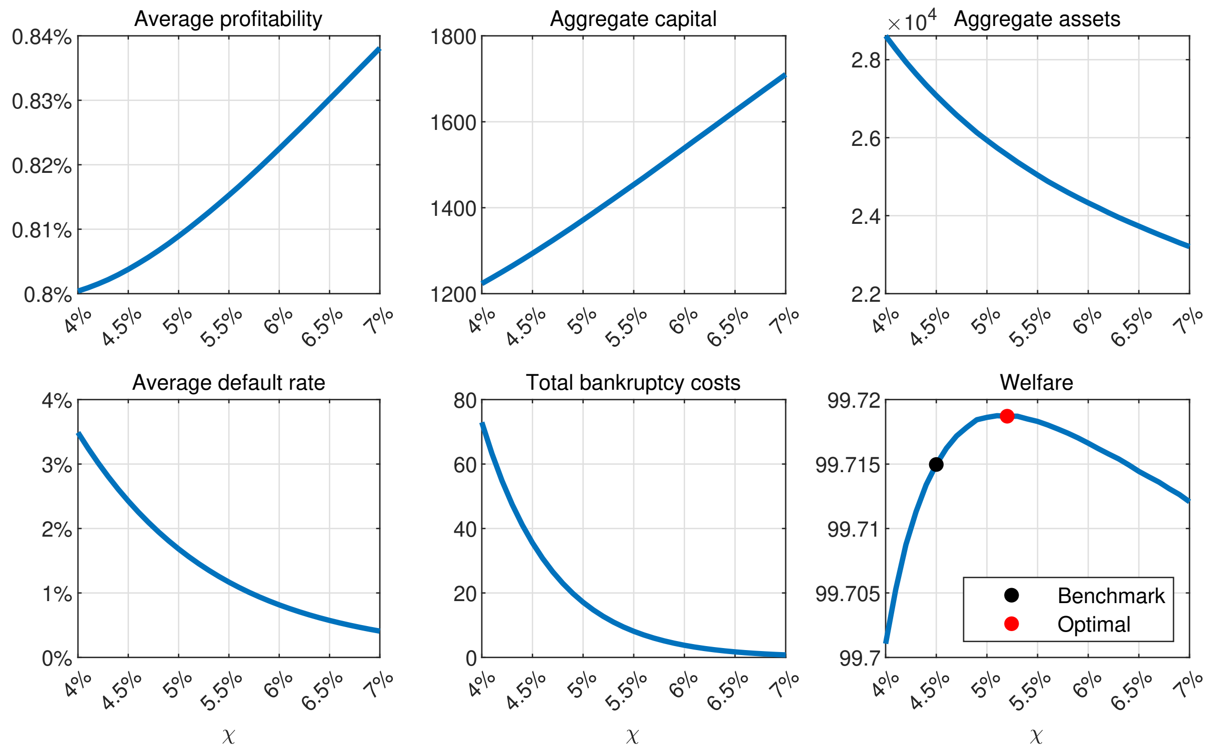

Figure 10: Aggregate outcomes as a function of the minimum capital-ratio requirement χ.

The transmission chain I am interested in is as follows. As regulation tightens, individual banks adjust their behavior. This leads to shifts in the dynamics of the banking sector, especially the size-distribution. In turn, the various macroeconomic aggregates, including output and bankruptcy costs, also adjust. This has implications for how much households eventually consume, and therefore their welfare. In Figure 10, I present a series of computations to show exactly how a change in regulation transmits through the model economy.

One, average profitability increases as χ increases (first panel). This has to do with a rightward shift in the distribution of banks, and the fact that larger banks are more efficient.

Second, the total amount of bank capital in the economy increases (second panel). This finding underscores that even when banks cannot raise capital externally, the banking sector as a whole responds to tighter regulation by accumulating more capital via retained earnings. This increase in aggregate capital acts as a counteracting force to regulation and underscores the *banking-dynamics channel* of regulation.

Third, the total amount of financial intermediation in the economy declines. This is because while aggregate capital increases, the fact that each bank is more constrained is the dominating force. Accordingly, aggregate assets of the banking sector (or total bank credit) declines (third panel). Relatedly, as banks become less leveraged, the average default rate among banks drops (fourth panel).

Total bankruptcy cost, which is formally given as $EL = PD \times EAD \times LGD$, also declines (fifth panel), although this is not obvious ex-ante.[23] On the one hand, banks' PDs decline as regulation tightens. On the other hand, as the distribution of banks shifts and there are fewer smaller-sized and more middle-sized banks, both the EAD and the LGD of the banking sector increase. The combined impact on bankruptcy cost is therefore ambiguous. Nonetheless, the simulations reveal that the former, i.e. PD effect, dominates. The decline in bankruptcy cost is an important channel via which tighter regulation produces welfare gains.

The welfare-improving and welfare-decreasing effects of tighter regulation, taken together, result in an inverted *U-shaped* response in aggregate welfare (sixth panel). The welfare maximising level of χ is around 5.2%, tighter relative to the benchmark of 4.5%. While not directly comparable, this result points in the same direction as that suggested in Begenau [2020], Admati and Hellwig [2014], Nguyen [2015] and Fender and Lewrick [2016].[24]

The gain in welfare from a tightening of regulation from 4.5% to 5.2% can be expressed in terms of consumption equivalence (CE), defined as the fractional increase ν in consump-

[23] PD = probability of default, EAD = exposure at default, LGD = loss given default.

[24] I take the Common Equity Tier-1 capital or core capital requirement in Basel III as the benchmark and derive the implications for optional regulation on that basis. This is because the G-SIB requirements – a focal point in this paper – are cast in terms of core capital. In doing so, I abstract away from any add-on capital requirements such as the counter-cyclical or capital conservation buffers. This helps explains why the optimal level in this paper is lower than other studies on capital regulation that consider the total capital ratio requirement as the benchmark.

tion that the household would receive should it live in the optimal regime forever. That is, if A denotes the benchmark regime and B denotes the optimal capital regulation regime, then ν is given as:

$$u\Big((1+\nu)C^A\Big)/(1-\beta) = u\Big(C^B\Big)/(1-\beta)$$

The value of ν in this case is 1.09%, suggesting that tighter regulation leads to a material improvement in the household's consumption.

6.2 Varying default losses and efficiency: $\Delta(s)$ and $\sigma(s)$

I begin by testing whether the main results are robust to the choice of specific functional forms for the mean and standard deviation of return on assets. In particular, if I consider $\theta(s) = \hat{\theta}_0 + \hat{\theta}_1 e^s/(1+e^s)$ and $\sigma(s) = \hat{\sigma}_0 - \hat{\sigma}_1 e^s/(1+e^s)$ where $\hat{\theta}_0 = 1.01, \hat{\theta}_1 = 0.01, \hat{\sigma}_0 = 0.03, \hat{\sigma}_1 = 0.01$, I continue to get a hump shaped welfare response and a similar value for the optimal regulation, i.e. 5.1%, as in the last panel in Figure 10.

I then consider a series of comparative statics. A higher loss given default (LGD) can strengthen the case for regulation. This is because when $\Delta(s)$ increases, the overall bankruptcy cost increase *ceteris paribus*. Meanwhile, reducing bankruptcy cost is the main channel via which regulation improves welfare. In line with this intuition, I find that as $\Delta(s)$ increases – either in terms of its level for all banks or in terms of the systemic loss component in the case of large banks – the optimal regulation becomes more stringent (see first two panels of Figure 11). Notably, the welfare gain in going from the benchmark to the optimal policy regime is greater when $\Delta(s)$ is larger. This can be seen by comparing the vertical distance between the black and red dots when $\Delta(s)$ is low versus high. The figure also illustrates that as χ increases beyond optimal levels, the welfare difference between alternative $\Delta(s)$ profiles becomes smaller. This is because as banks' default probabilities approach zero due to a higher χ, any change in the $\Delta(s)$ becomes less relevant.

The degree of efficiency can also affect optimal regulation. As the last panel of Figure

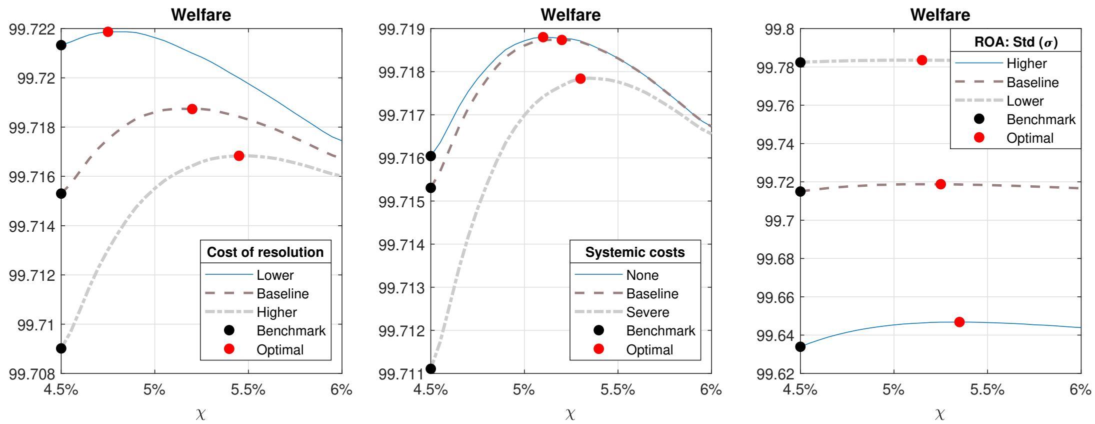

Figure 11: The impact of changes in the the profile of $\Delta(s)$ and $\sigma(s)$ on welfare and optimal regulation. The first panel considers variation in resolution losses, which has to do with the *level* of $\Delta(s)$ while keeping its non-linear profile intact i.e. the systemic loss increases with bank size beyond the 80th percentile and equals 23 percent of GDP for the largest bank. *Lower, baseline*, and *higher* correspond to resolution losses being 12%, 22%, and 32% of assets respectively. The second panel considers variations in the systemic loss associated with large bank defaults. *None* assumes no systemic losses. *Baseline* is the same as in the previous panel. *Severe* considers systemic losses that increases with size beyond the median, and equals 63 percent of GDP for the largest bank (also see Figure 4). The third panel studies the impact of increasing or decreasing the standard deviation of return on assets by a tenth.

11 shows, when the variance in return on assets increases, overall welfare declines and the optimal regulation becomes more stringent. The findings in this subsection underscore that both bankruptcy costs and efficiency are essential determinants of the optimal level of regulation.

6.3 Role of industry dynamics: $\mu(s)$

A natural question that arises is how important is the banking-dynamics channel – i.e. the endogenous response in the size-distribution $\mu(s)$ and entry-exit of banks – for welfare and policy implications. To assess this, we consider a counterfactual setting where the distribution of banks is kept *fixed* as in the benchmark equilibrium, and only individual banks' behaviours are allowed to respond to a change in χ. Figure 12 show how two key macroeconomic aggregates response to a higher χ in this setting. Because the distribution of banks remains fixed, and because for each given level of capital the corresponding bank is more constrained in its ability to invest in assets, aggregate assets decline by *more* as compared to the baseline case (first panel). Indeed, unlike in the baseline, now there is no countervailing force to regulation in the form of a rightward shift in the bank size distribution and the attendant *increase* in aggregate capital (recall second panel of Figure 10). As a result, in this setting, welfare declines precipitously as regulation tightens (second panel), whereas in the baseline setting, it followed an inverted U-shape.

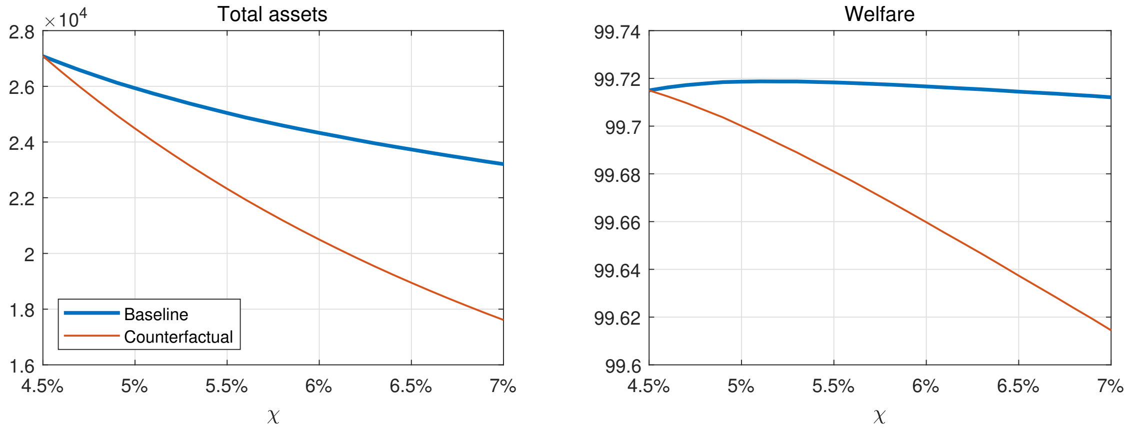

Figure 12: The panels contrast the response in total assets and welfare in the baseline case with a counterfactual case where the distribution of bank capital is held fixed as in the baseline economy.

7 Risk- and size-dependent regulation

Thus far I considered a regulatory regime where all banks face the same minimum capital-ratio requirement. However, regulation can be imposed differently. In particular, regulation could be a function of various bank specific characteristics such as its riskiness, systemic impact, or size. I consider these regulatory variants in the following subsections.

7.1 Equating the probability of default across banks

A well established regulatory approach is to impose risk-sensitive capital requirements, as in case of the Basel II framework. Basically, this entails a more stringent requirement for banks whose assets are more risky. A perfectly risk-sensitive requirement may, in fact, equate banks' probability of default (PD).

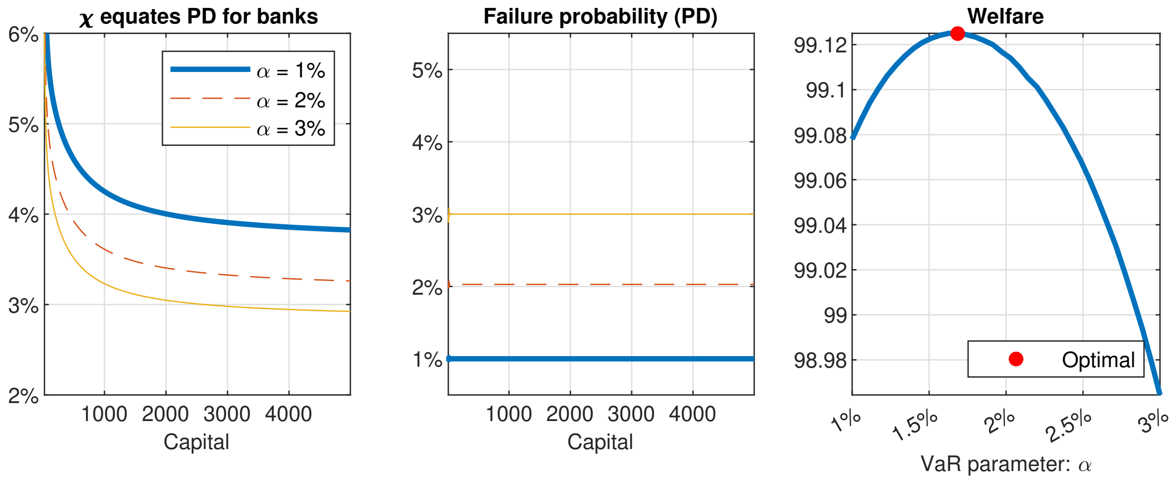

Figure 13: Risk-sensitive regulation that equates the probability of default (PD) across banks.

To assess the implications of such a requirement, I consider $\mathcal{X}(.)$ to be such that for any given bank size n, $Pr_n(\psi' < \psi^c(n)) = \alpha$ where $\psi^c(n)$ is the default cutoff, and the subscript Pr_n reflects the fact that the distribution of ψ' potentially depends on n. α can be thought of as a Value-at-Risk (VaR) parameter that governs the stringency of the requirement: lower value of α implies tighter regulation. As the first two panels in Figure 13 show, the capital requirement needed to equalise PD across banks is one that is less stringent for larger banks. This is because larger banks have an inherently more favorable risk-return profile, and thus can satisfy the same PD as a smaller bank while maintaining a lower capital-ratio (i.e. higher leverage). As regulatory stringency captured by α changes, welfare traces an inverted U-shaped profile, as in the previous analysis (third panel). However, the maximum welfare achieved in this regime in lower than that achieved in the previous analysis. The reason for this sub-optimal welfare result in this case is that while large and small banks have the same PD, the expected loss (EL) in case of large bank defaults is higher. Because EL is a key input to welfare in the economy, an unbalanced EL distribution leaves scope for welfare to be improved. By contrast, in the previous regime, where all banks are equally leveraged, larger banks end up having a smaller PD, which makes their EL relatively more comparable to that of smaller banks.

7.2 Equating expected losses across banks

Inspired by the insight from the previous analysis, I now consider the case where $\chi(.)$ is such that the expected loss (EL) posed by each bank, given as $PD \times EAD \times LGD$ where $EAD = s$ and $LGD = \Delta(s)$, is equalised. Although not explicitly stated as such, this is roughly the idea behind the Basel III G-SIB framework (see BCBS [2018]) that strives to mitigate systemic risks posed by larger banks by imposing a greater capital requirement on them. Indeed, the failure of (or distress among) large banks drove systemic losses during the Great Financial Crisis of 2008, leading to the adoption of the GSIB framework.

In contrast with the previous analysis, larger banks face a more stringent requirement in this regime, irrespective of the targeted level of EL (first panel of Figure 14). The kinks are noteworthy. They correspond to the point where systemic losses kick in. Relatedly, the slope of regulation is steeper beyond this point to compensate for the higher LGD for larger banks. The implied PDs and ELs are plotted in the second panel. By design, EL is the same across banks, but because $LGD \times EAD$ is greater for larger banks, this must be compensated by a smaller PD. Here again, the kinks in the PDs (which correspond to the kinks in regulation in the first panel) can be noted.

As regulatory stringency increases, welfare traces an inverted U-shaped profile, similar to all previous analyses. Interestingly, however, the maximum welfare achieved in this regime is the highest so far. The intuition for this result is that equating EL across banks takes into account the fact that all its sub-components (i.e. PD, LGD, EAD) depend on bank size and ultimately on bank size distribution.

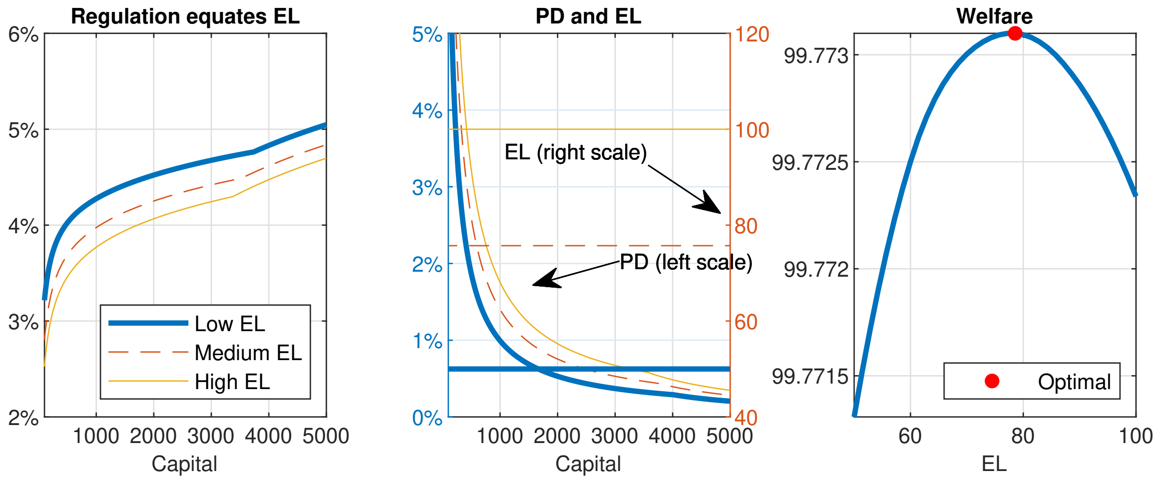

Figure 14: Regulation that equates expected losses (that is, EL = PD × LGD × EAD) across banks.

7.3 Size-dependent policy

The three alternative regulatory regimes considered so far equalise a specific metric across banks. These are (i) the non-risk-weighted capital-ratio (i.e. leverage), (ii) the probability of default PD, or (iii) the expected loss EL, respectively. The shortcoming of these rules, however, is that they still do not fully internalize the efficiency versus financial-stability trade-off.

To see this, consider the first regime studied above. In that regime, the capital requirement is the same for all banks, and does not take any account of differences across banks. In the second regime, a perfectly risk-sensitive requirement takes into account that large banks are more efficient and thus less likely to default (lower PD), but abstracts away from how EAD and LGD depend on bank size. The third regime improves upon the first two by taking PD, LGD, and EAD into account and equalising EL across banks. In doing so, the third regime acknowledges that larger banks are less likely to fail (ceteris paribus) but also more costly to resolve when in default. Yet, it is not obvious why equalising EL across banks should be the principal criteria behind setting χ. EL is only one determinant of aggregate welfare in the economy – the value of financial intermediation in the economy, underpinned by the aggregate assets of the banking sector, is another key component of overall welfare. This component also depends on how banks respond to regulation, and therefore, only optimising on the basis of EL falls short off fully optimising the efficiency versus financial-stability trade-off. In other words, equalising PD or EL across banks optimises *cost* or *risk* related aspects of the banking sector but does not take into account the *benefits* or value added it brings. To this end, in this section I consider a fully flexible size-dependent capital requirement and assess the welfare implications of such a regime.

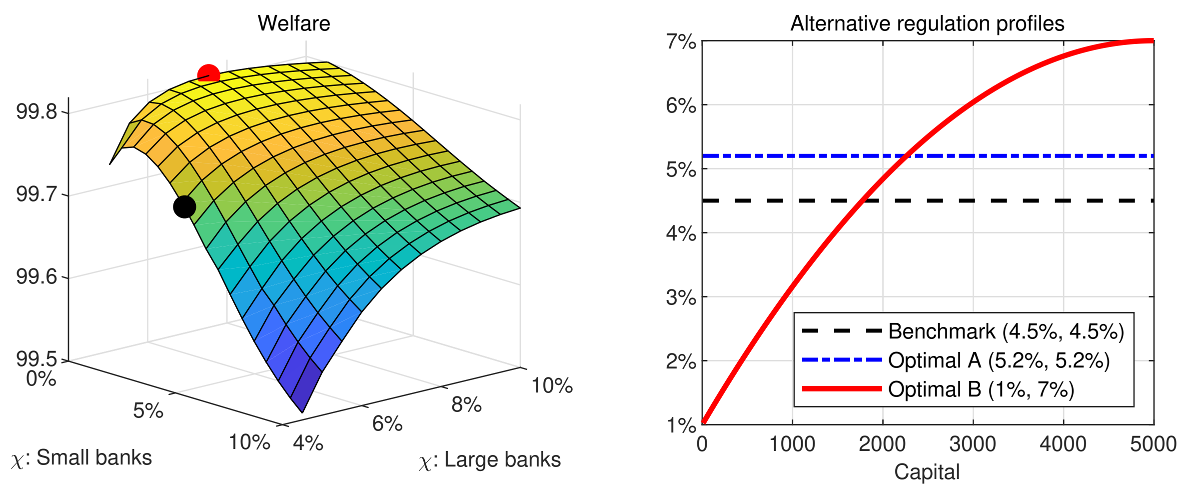

Figure 15: The first panel shows the welfare as a function of (χ_s, χ_l). Each point on the grid corresponds a specific regulation profile as per Equation (3). The second panel compares the benchmark and optimal regulation profiles. Optimal "A" corresponds to the optimal uniform regulation, while Optimal "B" corresponds to the optimal size-dependent regulation. The tuples in the legend indicate the values of χ_s and χ_l as described in Equation (3).

Formally, I assume that $\chi(n)$ can vary depending on the amount of capital a bank has.[25] To strike a balance between retaining a flexible specification and avoiding dimensionality issues, I consider $\chi(n)$ to have a quadratic form with three free parameters:

$$\chi(n) = \chi_0 + \chi_1 n + \chi_2 n^2$$

[25] Recall that size is measured by the amount of capital banks have. Alternatively, size could be measured by the amount of assets and regulation could be set on that basis. However, capital is a *state* variable for banks' decisions, and banks' assets are uniquely determined by the amount of capital they have. This means that in this model regulation that varies across banks on the basis of their assets would be equivalent to the one that varies on the basis of their capital.

To further discipline the optimisation problem of the regulator striving for the welfare maximising $\chi(n)$ profile, I restrict attention to $\chi(n) \in [0\%, 100\%]$. This ensures that the minimum requirement does not diverge to absurd values. I also assume the following limiting condition, $\lim_{n\to\overline{n}} \chi'(n) = 0$, to ensure that as bank size increases regulation stabilises at a certain level, which could even be a 100% capital requirement.[26] These conditions reduce the number of free parameters in $\chi(n)$ from three to two. In turn it allows $\chi(n)$ to be expressed in terms of χ_s and χ_l, the minimum requirement applicable to the smallest and the largest banks (i.e. $n = \overline{n}$) respectively:

$$\chi(n) = \Big(\chi_s - \chi_l\Big)\Big(n/\overline{n}\Big)^2 - 2\Big(\chi_s - \chi_l\Big)\Big(n/\overline{n}\Big) + \chi_s \tag{3}$$

The welfare profile as a function of (χ_s, χ_l) is hump-shaped as shown in the first panel of Figure 15. This is similar in spirit to the inverted U-shaped profile in the previous regimes. The welfare maximizing regulation profile is shown in right-hand panel. Compared to the benchmark of 4.5%, and the optimal uniform (i.e. independent of size) regulation of 5.2%, the optimal size-dependent regulation features a relatively stringent requirement of around 7% for the largest banks. Even though the rationale for regulation and underlying mechanisms are different, this result points in the same direction as Dávila and Walther [2020]. It also supports the spirit of the G-SIB framework in terms of imposing tighter regulation on the larger banks. In fact, quantitatively, the model implied requirement of 7% is close to what the G-SIB framework has set in place for the largest G-SIB (namely J. P. Morgan, which faces a surcharge of 2.5% on top of the baseline requirement of 4.5%). That said, in contrast to the G-SIB framework, the analysis suggests a more liberal requirement of close to 1% in the case of smaller banks.

The economic rationale for such a regime may be understood as follows. On the one

[26] I consider this parametric form for χ for tractability. Otherwise, the space of functions over which to optimise can become *high* dimensional. For instance, in principle, the $\chi(n)$ can be a step function – as in the case of the G-SIB framework – but this requires many more free parameters to optimise on and is computationally challenging.

hand, this regime allows smaller banks to assume higher leverage, which has the benefit that they can potentially grow faster and rapidly benefit from scale economies. Obviously, this means that small banks default more often, but small bank defaults are also socially less costly. On the other hand, once banks become large and their default more costly, this regime limits the expected loss (EL) posed by them by lowering their default rate via a higher capital requirement.

Qualitatively, this regime is comparable to the one that equates EL across banks in the sense that the optimal regulation is tighter for larger banks (recall first panel of Figure 14). However, a flexible size-dependent regime achieves higher welfare still. While this result is mathematically obvious since the regulator is less constrained in this regime, the main insight is about where the welfare gain comes from. As discussed above, the welfare gain stems from the fact that a policy rule that only equates EL across banks fails to account that both efficiency and financial-stability depend on bank size and its distribution. Quantitatively, the welfare gains from adopting a flexible regime are substantial, at around 12% in consumption equivalence terms.

Varying the loss-given-default $\Delta(s)$ and the standard deviation of return on assets $\sigma(s)$ – in particular how they depend on bank size s – confirms the insights presented earlier. The first panel of Figure 16 shows that as LGD increases, the optimal level of regulation shifts upwards. In particular, the regulatory profile becomes 'steeper' when the systemic loss component, i.e. the LGD for larger banks, increases. The second panel documents that as the standard deviation of return on assets $\sigma(s)$ becomes smaller, the the optimal level of regulation shifts downwards. In particular, the regulatory profile becomes 'flatter' when the riskiness of larger banks declines. Overall, how size-dependent optimal regulation should be depends on how size-dependent bankruptcy costs and bank efficiency are.

All that said, it may be noted that a more flexible policy rule can be more difficult to implement in practice. Simpler rules that equate some tangible metric across banks can be easier to state, implement, and ensure compliance to as compared to a rule that

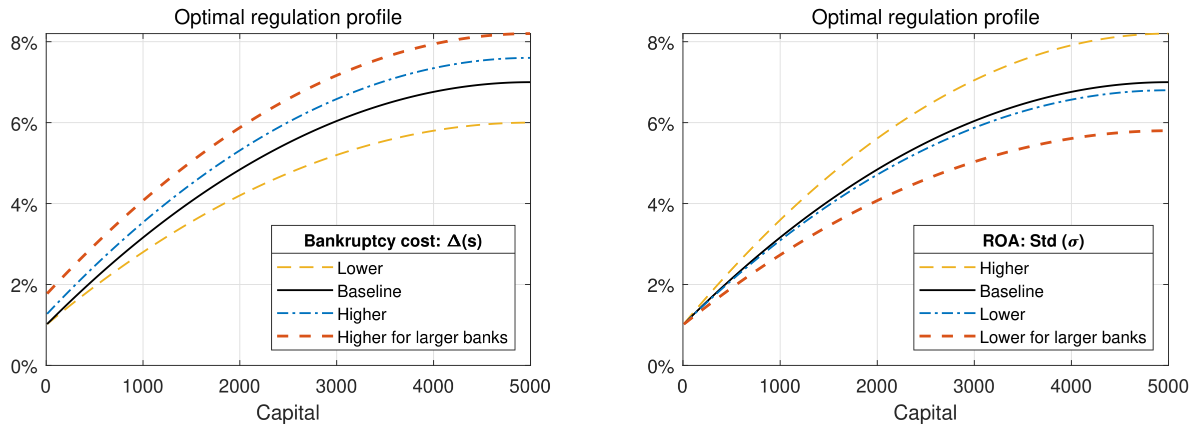

Figure 16: *First panel*: Optimal size-dependent regulatory profiles depending on the loss-given-default profile. The *baseline* regime is the one used throughout this paper, i.e. resolution losses equal 22% of assets for all banks and the systemic loss is applicable to banks beyond the 80th percentile, increasing with bank size equalling 23% of GDP for the largest bank. The *Lower* and *Higher* regimes consider resolution losses to be 12% and 32% of assets respectively. The *Higher for larger banks* regime increases the systemic losses to thrice the GDP. *Second panel*: Optimal size-dependent regulatory profiles depending on the variance of return on assets. The *Higher* and *Lower* regimes increase or decrease respectively the standard deviation of return on assets by a tenth relative to the baseline. The *Lower for larger banks* regime further reduces the standard deviation of return on assets for the larger banks: $\sigma(s) = 0.9(\sigma_0 + \sigma_1/(1+2s))$.

imposes bank-specific requirements based on a complex formula.

8 Endogenous returns on assets and mass of banks

Thus far I have assumed that the return on banks' assets is given exogenously, and that the mass of banks is fixed at unity. While these assumptions make the model more tractable, they abstract away from two potential channels through which regulation can transmit. In this section, I relax these assumptions and reassess the regulatory and welfare implications.

8.1 Response of returns on assets to regulation

I first endogenize the return on assets. To this end, I assume that the banking sector as a whole faces a downward sloping demand for bank credit. That is, when banks collectively invest in more assets, the return on any individual bank's investments is lower. This means that as regulation tightens and alters the level of banks' aggregate investment, the

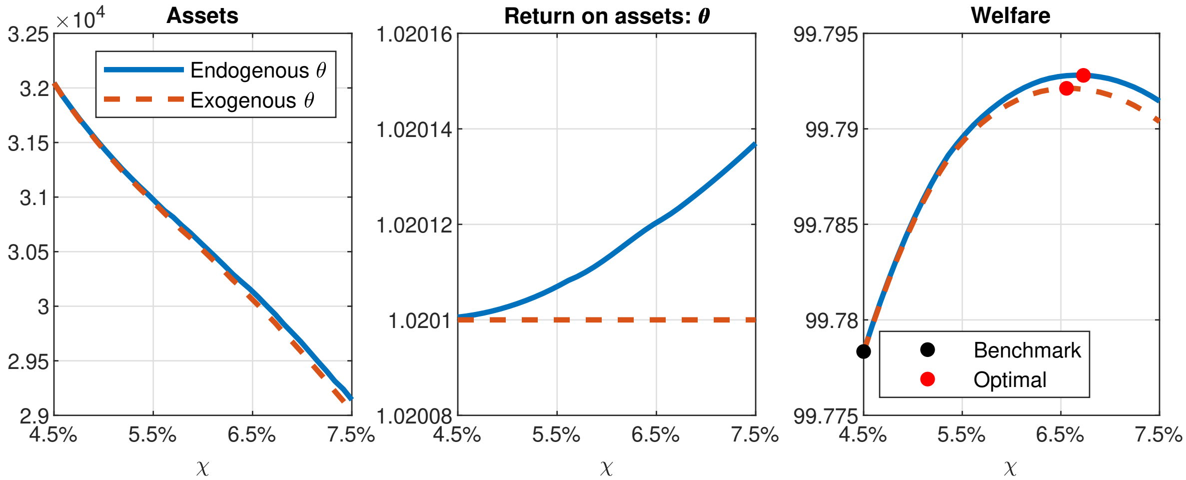

Figure 17: Response to changes in the minimum capital-ratio requirement when return on assets is endogenous.

return on assets also changes. This then triggers a *second-round effect* on individual banks' behaviors, which is new relative to the channel in the benchmark model. Ultimately, this has implications for aggregate welfare and optimal regulation.

To analyse the role of this channel, I assume that θ, the expected return on any individual bank's assets, is a decreasing function of aggregate bank assets S in the economy:[27]

$$\theta(S) = \rho_0 + \frac{\rho_1}{(1 + S/S_{benchmark})},$$

where I assume that $\theta(S)$ is such that $\theta(S_{benchmark}) = \theta_0$ where recall that θ_0 (Table 1) is the component of θ that does not depend on an individual bank's size. For tractability of studying this new channel, I switch off bank-level economies of scale, so that $\theta_1 = 0$. One parameterization that is consistent with this condition is $\rho_0 = 1.0191, \rho_1 = 0.002$. With that, I assess the implications of an increase in the minimum capital-ratio requirement

[27]The reduced form approach to modeling the relation between θ and S in this paper is similar to the one adopted in the seminal Monte-Klein model (Klein [1971]; Monti [1972]), and also more recently in Liu [2019]. That said, this relation can be micro-founded by assuming a representative firm that seeks bank funding and exhibits decreasing returns to capital, like in Gertler and Kiyotaki [2010] for example.

above its benchmark value of 4.5%.[28]

As shown in the first panel of Figure 17, as χ increases, total assets of the banking sector declines (blue line). But compared to the benchmark economy, the decline is smaller (compare blue line with red line). This is because as aggregate assets decline, return on assets increases (second panel), which improves the ability of banks to generate earnings and build capital. This further pushes the bank size-distribution rightwards (in a stochastic dominance sense). The extended model, therefore, embeds an additional countervailing force in response to tighter regulation. In turn, this allows regulation to push harder as the welfare cost of tighter regulation – i.e. banks becoming constrained and being able to intermediate less – is lower in a *ceteris paribus* sense. Indeed, as the third panel shows, optimal χ as well as the maximised welfare in the endogenous θ case are higher than in the exogenous θ case.[29]

8.2 Response of mass of banks to regulation

In the benchmark economy, recall that insolvent banks re-enter the industry upon receiving a random amount of seed capital. This captures the spirit of bank entry-exit in practice wherein insolvent banks are typically merged with incumbent banks after some capital injection by the acquirer, but abstracts away from the possibility that an additional mass of banks may enter the industry each period depending on how profitable it is to do so. To allow for this profitability-dependent entry in the model, and therefore to have an endogenously determined equilibrium mass of banks, I consider the following extension of the benchmark model.

[28]The algorithm to solve the model for a given level of regulation in this case is more involved as compared to the one in the benchmark. For a guessed starting value of return on assets, θ_{guess}, I solve the model and compute the corresponding S. Then, I compute the $\theta_{implied}$ implied by the S according to the assumed functional relation between them. Finally, I adjust the guess as follows: $\theta_{guess} \rightarrow (1-\omega)\theta_{guess} + \omega\theta_{implied}$ where ω is the adjustment weight that I set to 0.33 (to keep the updating process stable).

[29]Note that the exogenous θ case is not identical or directly comparable to the benchmark regime as θ_1 is assumed to be zero in this case in order to focus on the role that an endogenous θ_0 plays.

I assume that there is a mass of potential entrants whose opportunity cost of entering the banking sector is randomly distributed as per $\mathcal{F}_e(.;\theta_e,\sigma_e)$, where $\mathcal{F}_e$ is a cumulative normal distribution with mean θ_e and standard deviation σ_e. The opportunity cost can be thought of as the value of an investment project that the entrant has access to – the so-called *outside option*. This means that a potential entrant enters the banking industry if the expected present discounted value of entering:

$$EV_e = \int V(n_e)dG(n_e)$$

is greater than their opportunity cost (as also in Hopenhayn [1992]). So, when EV_e increases, there are more entrants. The relationship between EV_e and entrant mass can then be expressed as follows:

$$M(EV_e) = \alpha_e \mathcal{F}_e(EV_e;\theta_e,\sigma_e),$$

where M is the mass of entrants and α_e is a scaling factor. What this specification means is that as EV_e increases relative to θ_e, the mass of entrants increases asymptotically to a maximum of α_e. To facilitate a comparison with the benchmark economy, I impose the condition that $M(EV_e^{benchmark}) = M_{benchmark}$, and consider the following parameterization that is consistent with the above specification: $\alpha_e = 1.01 * M_{benchmark}$, $\theta_e = EV_e^{benchmark}/1.05$, and $\sigma_e = 0.02 * EV_e^{benchmark}$.

Next, I assess the implications of a change in the minimum capital-ratio requirement from its benchmark value of 4.5%.[30] As regulation tightens, banks become more constrained, and the expected present discounted value of all banks declines. As a result, the expected value of entry into the banking sector also declines, and fewer entrants enter the

[30]The solution algorithm in this case is similar to the one where return on assets is endogenous, except that in each iteration, there are three additional steps. For a given value of θ_{guess}, first EV^e is computed. Then the mass of entrants M is backed out. Correspondingly, the mass of incumbents is obtained, which in turn is used to compute the model implied value of S and eventually $\theta_{implied}$.

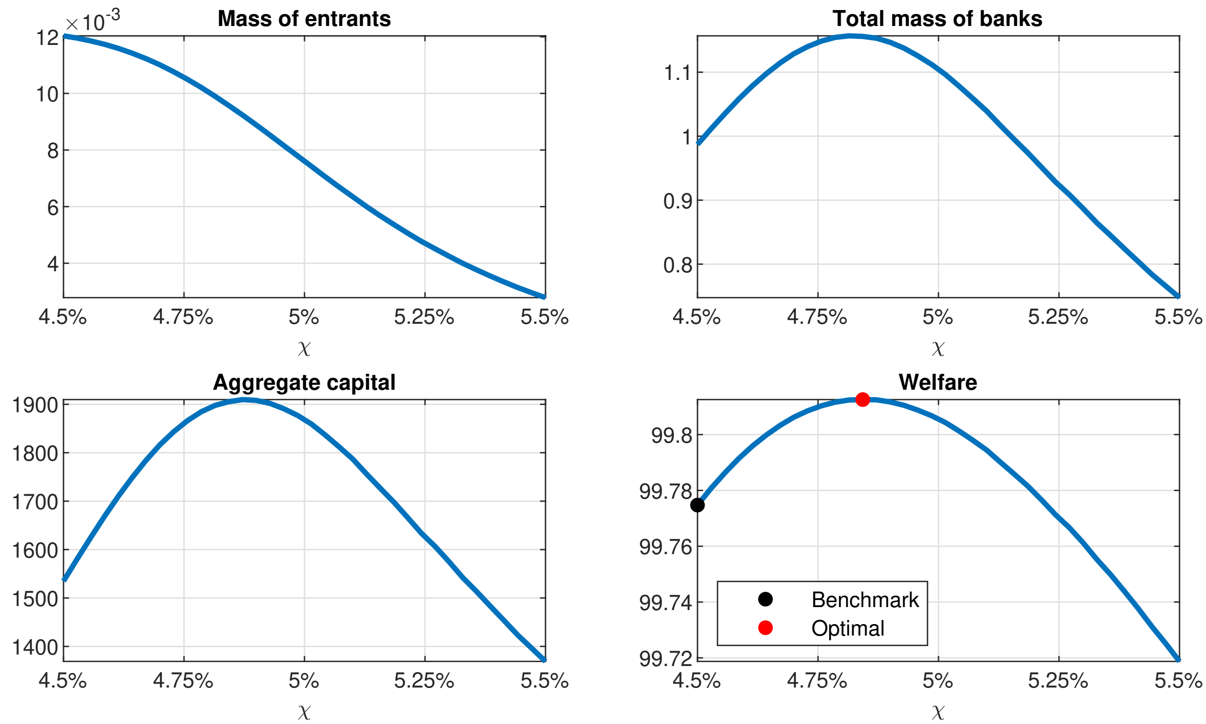

Figure 18: Response to changes in the minimum capital-ratio requirement when the mass of banks is endogenous.

industry (first panel of Figure 18). The total mass of incumbent banks, however, follows an inverted U-shaped pattern (second panel). This non-monotonic response is because the mass of incumbent banks depends on how fast the mass of entrants declines *relative* to the drop in incumbent banks' average default rate.

To illustrate this point via an example, let the mass of entrants be M, the mass of incumbents be M^I, and the default probability be p. In the stationary competitive equilibrium, these quantities must satisfy the relation $p = M/M^I$. When regulation tightens, p drops since banks are better capitalised, while M drops since expected value of entry is lower. Then, if p drops by a lot more (less) relative to M, mass of incumbents would increase (decrease).

Relatedly, it is useful to note that initially the decline in default rate is rapid, driving up the mass of incumbents, while once there is a more rapid decline in expected value of entry, entry also drops substantially, which brings down the equilibrium mass of banks. Aggregate capital, which is closely related to the mass of incumbent banks, also follows an inverted U-shaped pattern (third panel). This is in contrast to the benchmark where aggregate capital increased monotonically for the range of χ considered (recall Figure 10).

The policy implication in this case is that the optimal regulation (fourth panel) is less stringent as compared to the optimal in the benchmark economy (where it was around 5.2%). Indeed, by constraining banks, tighter regulation creates disincentives for potential entrants to enter the banking sector, and makes regulation more costly than in the benchmark economy.

9 Conclusion

This paper in concerned with an efficiency versus financial-stability trade-off in banking – the fact that larger banks may be more efficient but their default can be socially more costly. It strives to understand how banks should be organized given this trade-off – few large or many small ones – and how capital regulation can be used to balance this trade-off and implement the desired distribution of bank sizes. The goal is both positive and normative analysis of capital regulation.

To achieve its goal, the paper develops a tractable general equilibrium macroeconomic model of a heterogeneous banking sector with endogenous size-distribution and entry-exit. Two core aspects of the model are that the organisation of the banking sector features a non-trivial response to regulation, and that there is an explicit welfare rationale for regulation, namely the bank default externality.

The paper shows that banking industry dynamics is an important channel through which capital regulation operates. While individual banks are obviously affected by regulation, regulation also shapes the overall dynamics of the banking sector, especially the size-distribution of banks. In turn, this has aggregate implications that do not necessarily go in the same direction and thus pose a trade-off for the regulator. For instance, tighter regulation leads to a rightward shift in the size-distribution of banks and at the same time a less heavy right-tail. While this leads to an increase in the aggregate capital stock, the impact on aggregate efficiency and expected default losses is ambiguous.

Given this ambiguity, the paper uses a series of counterfactual experiments to determine the optimal regulation. It shows that a capital requirement regime that equates leverage, default rate, or expected default losses across banks falls short of balancing the efficiency versus financial-stability trade-off. This is because these regimes focus on minimising the costs and risks posed by the banking sector, but fail to take account of how the value-added that banks bring – including via their efficiency – responds to regulation. The paper shows that the optimal regulation should, therefore, be more flexible and vary with banks' sizes.

The paper lends support to the idea of imposing tighter regulation on larger banks, as in case of the G-SIB framework, but also stresses that regulation can do better by easing the requirement for the smallest banks whose failures are less costly. At the same time, the paper highlights the material welfare implications of taking differences in bank efficiency seriously. This can be increasingly relevant for policy design today as technology companies and start-ups tend to rapidly gain market share in the financial sector on the back of data-driven scale economies. The tractable model developed in this paper can facilitate the analysis of these issues in subsequent research.

References

Admati, A. and M. Hellwig (2014). *The Bankers' New Clothes: What's Wrong with Banking and What to Do about It.* Princeton University Press.

Adrian, T. and N. Boyarchenko (2012). Intermediary leverage cycles and financial stability.

Becker Friedman Institute for Research in Economics Working Paper (2012-010).

Albuquerque, R. and H. A. Hopenhayn (2004). Optimal lending contracts and firm dynamics. *The Review of Economic Studies 71*(2), 285–315.

Atkinson, T., D. Luttrell, and H. Rosenblum (2013). How bad was it? the costs and consequences of the 2007–09 financial crisis. *Staff Papers* (Jul).

Baele, L., O. De Jonghe, and R. Vander Vennet (2007). Does the stock market value bank diversification? *Journal of Banking & Finance 31*(7), 1999–2023.

BCBS (2010). An assessment of the long-term economic impact of stronger capital and liquidity requirements. *A report by the Basel Committee on Banking Supervision*.

BCBS (2018). Global systemically important banks (G-SIBs): Revised assessment methodology and the higher loss absorbency requirement. *Basel Committee for Banking Supervision (BCBS)*.

Beccalli, E., M. Anolli, and G. Borello (2015). Are European banks too big? Evidence on economies of scale. *Journal of Banking & Finance 58*, 232–246.

Begenau, J. (2020). Capital requirements, risk choice, and liquidity provision in a business-cycle model. *Journal of Financial Economics 136*(2), 355–378.

Bellifemine, M., R. Jamilov, and T. Monacelli (2022). Hbank: Monetary policy with heterogeneous banks. *CEPR Discussion Paper DP17129*.

Bernanke, B. S., M. Gertler, and S. Gilchrist (1999). The financial accelerator in a quantitative business cycle framework. *Handbook of macroeconomics 1*, 1341–1393.

Bianchi, J. and S. Bigio (2022). Banks, liquidity management, and monetary policy. *Econometrica 90*(1), 391–454.

Boissay, F., F. Collard, and F. Smets (2016). Booms and banking crises. *Journal of Political Economy 124*(2), 489–538.

Brunnermeier, M. K. and Y. Sannikov (2014). A macroeconomic model with a financial sector. *American Economic Review.*

Caballero, R. J. and A. Simsek (2013). Fire sales in a model of complexity. *The Journal of Finance 68*(6), 2549–2587.

Christiano, L. and D. Ikeda (2013). Leverage restrictions in a business cycle model. *NBER Working Paper 18688*.

Christiano, L. and D. Ikeda (2016). Bank leverage and social welfare. *American Economic Review Papers and Proceedings 106*(5), 560–64.

Clementi, G. L. and H. A. Hopenhayn (2006). A Theory of financing constraints and firm dynamics. *The Quarterly Journal of Economics 121*(1), 229–265.

Cooley, T. F. and V. Quadrini (2001). Financial markets and firm dynamics. *American Economic Review 91*(5), 1286–1310.

Corbae, D. and P. D'Erasmo (2021). Capital buffers in a quantitative model of banking industry dynamics. *Econometrica 89*(6), 2975–3023.

Corbae, D. and P. D'Erasmo (2010). A quantitative model of banking industry dynamics. *Society for Economic Dynamics Meeting Papers 268*.

Corbae, D. and P. D'Erasmo (2020). Rising bank concentration. *Journal of Economic Dynamics and Control 115*, 1038–77.

Dávila, E. and A. Walther (2020). Does size matter? Bailouts with large and small banks. *Journal of Financial Economics 136*(1), 1–22.

De Nicolo, G., A. Gamba, and M. Lucchetta (2014). Microprudential regulation in a dynamic model of banking. *The Review of Financial Studies 27*(7), 2097–2138.

Fender, I. and U. Lewrick (2016). Adding it all up: The macroeconomic impact of Basel III and outstanding reform issues. *BIS Working Papers 591*.

Flannery, M., B. Hirtle, and A. Kovner (2017). Evaluating the information in the Federal Reserve stress tests. *Journal of Financial Intermediation 29*, 1–18.

Frost, J., L. Gambacorta, Y. Huang, H. S. Shin, and P. Zbinden (2019). Bigtech and the changing structure of financial intermediation. *Economic policy 34*(100), 761–799.

Gambacorta, L. and A. A. van Rixtel (2013). Structural bank regulation initiatives: Approaches and implications. *BIS working paper 412*.

Gârleanu, N., S. Panageas, and J. Yu (2015). Financial entanglement: A theory of incomplete integration, leverage, crashes, and contagion. *American Economic Review 105*(7), 1979–2010.

Gertler, M. and N. Kiyotaki (2010). Financial intermediation and credit policy in business cycle analysis. *Handbook of Monetary Economics 3*, 547.

Goel, T., U. Lewrick, and N. Tarashev (2020). Bank capital allocation under multiple constraints. *Journal of Financial Intermediation*.

Hopenhayn, H. and E. Prescott (1992). Stochastic monotonicity and stationary distributions for dynamic economies. *Econometrica 60*(6), 1387–406.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica 60*(5), 1127–1150.

Hughes, J. P., J. Jagtiani, L. J. Mester, and C.-G. Moon (2019). Does scale matter in community bank performance? Evidence obtained by applying several new measures of performance. *Journal of Banking & Finance 106*, 471–499.

Hughes, J. P. and L. J. Mester (2013). Who said large banks do not experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation 22*(4), 559–585.

Jamilov, R. (2021). A macroeconomic model with heterogeneous banks. *SSRN Working Paper 3732168*.

Jamilov, R. and T. Monacelli (2021). Bewley banks. *SSRN Working Paper 3732172*.

Janicki, H. and E. S. Prescott (2006). Changes in the size distribution of us banks: 1960-2005. *FRB Richmond Economic Quarterly 92*(4), 291–316.

Jensen, M. C. and W. H. Meckling (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics 3*(4), 305–360.

Jermann, U. and V. Quadrini (2012). Macroeconomic effects of financial shocks. *American Economic Review 102*(1), 238–71.

Kang, A., R. Lowery, and M. Wardlaw (2015). The costs of closing failed banks: A structural estimation of regulatory incentives. *The Review of Financial Studies 28*(4), 1060–1102.

Kara, G. I. and S. M. Ozsoy (2020). Bank regulation under fire sale externalities. *The Review of Financial Studies 33*(6), 2554–2584.

Kareken, J. H. and N. Wallace (1978). Deposit insurance and bank regulation: A partial-equilibrium exposition. *Journal of Business 51*(3), 413–438.

Kihlstrom, R. E. and J.-J. Laffont (1979). A general equilibrium entrepreneurial theory of firm formation based on risk aversion. *Journal of Political Economy 87*(4), 719–748.

Klein, M. A. (1971). A theory of the banking firm. *Journal of Money, Credit and Banking 3*(2), 205–218.

Krusell, P. and J. Smith, Anthony A (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy 106*(5), 867–896.

Laeven, L. and F. Valencia (2013). Systemic banking crises database. *IMF Economic Review 61*(2), 225–270.

Liu, K. (2019). The impact of the dodd-frank act on small us banks. *SSRN Working Paper 3419586*.

Malherbe, F. (2020, July). Optimal capital requirements over the business and financial cycles. *American Economic Journal: Macroeconomics 12*(3), 139–74.

Monti, M. (1972). *Deposit, credit and interest rate determination under alternative bank objective function*. Elsevier.

Muller, C. (2022). Capital requirements, market structure, and heterogeneous banks. *IWH Discussion Papers 15/2022*.

Myers, S. C. and N. S. Majluf (1984). Corporate financing and investment decisions when firms have informationthat investors do not have. Report 0898-2937, National Bureau of Economic Research.

Nguyen, T. T. (2015). Bank capital requirements: A quantitative analysis. *SSRN Working Paper 2356043*.

Passmore, W. and A. H. von Hafften (2019). Are Basel's capital surcharges for global systemically important banks too small? *International Journal of Central Banking 15*(1), 107 – 157.

Santos, J. A. C. (2001). Bank capital regulation in contemporary banking theory: A review of the literature. *Financial Markets, Institutions & Instruments 10*(2), 41–84.

Shleifer, A. and R. Vishny (2011). Fire sales in finance and macroeconomics. *Journal of Economic Perspectives 25*(1), 29–48.

Stokey, N. L. and J. Lucas, Robert E (1989). *Recursive methods in economic dynamics.* Harvard University Press.

Van den Heuvel, S. J. (2008). The welfare cost of bank capital requirements. *Journal of Monetary Economics 55*(2), 298–320.

Van Lelyveld, I. and K. Knot (2009). Do financial conglomerates create or destroy value? Evidence for the EU. *Journal of Banking & Finance 33*(12), 2312–2321.

Wang, Y., T. M. Whited, Y. Wu, and K. Xiao (2021). Bank market power and monetary policy transmission: Evidence from a structural estimation. *The Journal of Finance 77*(4), 2093–2141.

Wheelock, D. C. and P. W. Wilson (2012). Do large banks have lower costs? New estimates of returns to scale for US banks. *Journal of Money, Credit and Banking 44*(1), 171–199.

Wheelock, D. C. and P. W. Wilson (2018). The evolution of scale economies in US banking. *Journal of Applied Econometrics 33*(1), 16–28.

Whited, T. M., Y. Wu, and K. Xiao (2021). Low interest rates and risk incentives for banks with market power. *Journal of Monetary Economics 121*, 155–174.

Zhu, H. (2008). Capital regulation and banks financial decisions. *International Journal of Central Banking 4*(1), 165 – 212.

Appendices

A Characteristics of the bank's problem

Proof. To keep the proof tractable and analytically feasible, I assume that deposit premium $t = 0$, default size cutoff $\tau = 0$, and that the capital constraint is not sensitive to size or riskiness of a bank, i.e. $\chi = (n-e)/s$. With these simplifications, I note that $s = (n-e)/\chi$, $d = (n-e)/(1/\chi - 1)$, $\psi_c = R(1-\chi)$, and $n' = [\psi'/\chi - R(1/\chi - 1)](n-e)$. The bank's problem can thus be written in terms of a single decision variable e as follows:

$$V(n) = \max_{e} \left(\mathcal{H}(e) + \beta \int_{R(1-\chi)} V\Big(\big[\psi'/\chi - R(1/\chi - 1)\big](n-e)\Big) f(\psi';\theta,\sigma) d\psi'\right)$$

I follow the strategy in Stokey and Lucas [1989] to prove that the Bellman operator underpinning the above equation has a unique fixed point. First note that the payoff $\mathcal{H}(e)$ is bounded since $0 \leq e \leq n$. Next, I show that the Blackwell Conditions are satisfied. Let $\mathcal{C}$ be the class of continuous functions on the non-negative real line $\mathcal{R}_+$, and define the Bellman operator $\mathcal{Q}$ for an arbitrary value function $q \in \mathcal{C}$ as:

$$\mathcal{Q}(q(n)) = \max_{e} \left(\mathcal{H}(e) + \beta \int_{R(1-\chi)} q\Big(\big[\psi'/\chi - R(1/\chi - 1)\big](n-e)\Big) f(\psi';\theta,\sigma) d\psi'\right)$$

The first Blackwell Condition has to do with the *Monotonicity* of the Bellman operator $\mathcal{Q}$, that is:

$$\forall\, l, q \in \mathcal{C} \text{ s.t. } l(n) \leq q(n)\ \forall n \in \mathcal{R}_+ \implies \mathcal{Q}(l(n)) \leq \mathcal{Q}(q(n))\ \forall n \in \mathcal{R}_+$$

Let e^l and e^q be the maximands of $\mathcal{Q}(l(n))$ and $\mathcal{Q}(q(n))$ respectively. Crucially, e^l is a feasible choice for the Bellman operator on q at n. This implies that:

$$\mathcal{Q}(l(n))\Big|_{\text{at } e^l} \leq \mathcal{Q}(q(n))\Big|_{\text{at } e^l} \leq \mathcal{Q}(q(n))\Big|_{\text{at } e^q}$$

where the $|_{\text{at } e}$ notation stands for the computation of the Bellman operator at e. The first inequality follows simply from the fact that $l(n) \leq q(n)$, while the second one follows from the fact that e_q maximises $\mathcal{Q}(q(n))$. The second Blackwell Condition has to do with the *Discounting* property of $\mathcal{Q}$:

$$\exists \Delta \in (0,1) \text{ s.t. } \forall q \in \mathcal{C},\ \forall a \geq 0,\ \forall n \in \mathcal{N} \implies \mathcal{Q}(q(n)+a) \leq \mathcal{Q}(q(n)) + \Delta a$$

To show this, consider:

$$\mathcal{Q}(q(n)+a) = \max_e \quad \left(\mathcal{H}(e) + \beta \int_{R(1-\chi)} \Big(q(n') + a\Big) f(\psi';\theta,\sigma)d\psi'\right)$$

$$\leq \underbrace{\max_e \left(\mathcal{H}(e) + \beta \int_{R(1-\chi)} q(n')f(\psi';\theta,\sigma)d\psi'\right)}_{\mathcal{Q}(q(n))} + \max_e \beta a \underbrace{\int_{R(1-\chi)} f(\psi';\theta,\sigma)d\psi'}_{\text{probability of solvency: } 0<\zeta<1}$$

The inequality follows from the fact that the maximized sum of two or more functions cannot exceed the sum of the maximized values of those functions. In the second maximization, $e = 0$ is the optimal choice since it results in the smaller variance in ψ' and thus a higher probability $\zeta \in (0,1)$ of remaining solvent. In turn, this implies that the second term is equal to $\beta\zeta a$. With $\Delta = \beta\zeta$, this verifies the second Blackwell condition. Thus, the Bellman operator has a unique fixed point, say V, which corresponds to the bank's value and policy functions.

To show that V is increasing, let $n_1 < n_2$ be two potential states for the bank's problem, and let e_1 be the corresponding policy choice when state is n_1. Then choosing

$e_2 = n_2 - n_1 + e_1 > e_1$ effectively keeps the *post dividend* capital in state n_2 the same as that in state n_1. As a result, the policy choices namely s and d and endogenous variables n' and ψ_c are also the same across the two states since all of them depend on post dividend capital. Finally, since a feasible dividend choice in case of state n_2, i.e. e_2 achieves strictly greater payoff relative to n_1 ($\mathcal{H}$ is an increasing function), the maximized value at n_2 i.e. $V(n_2)$ is greater than $V(n_1)$. ■

B Leverage and default

Proof. The default probability of a bank with post dividend capital $\hat{n}$ can be written as:

$$p(\hat{n}) = \int^{\psi^c} f(\psi'; \theta(s), \sigma(s)) d\psi' = F(\psi_c; \theta(s), \sigma(s))$$

where $F(.)$ is the cumulative distribution function of ψ'. Under the simplifying assumptions considered for this proposition, the default cutoff is the same for all banks and is given as $\psi^c = R(1 - \chi)$. The dependence of p on $\hat{n}$, therefore, results from the fact that $\theta(s)$ increases and $\sigma(s)$ decreases as $\hat{n}$ increases. This is because a larger $\hat{n}$ leads to a larger $s = \hat{n}/\chi$, and that $\theta'(s) > 0$ and $\sigma'(s) < 0$. Therefore, $p'(\hat{n}) < 0$. ■

C Homogeneity of distribution evolution operator $\mathcal{T}$

Proof. Let μ_M be the stationary distribution corresponding to $M : \mu_M = T(\mu_M, M)$. Next, define a transition function $W(n, N)$ which denotes the probability that a bank of size n on date t evolves into a bank of size in the range $[\tau, N]$ on date $t+1$. Note that this implicitly means that defaults are excluded when accounting for the transition. Then,

$$\mu_M(N) = M \int_0^N dG(n_e) + \int_\tau W(n, N) d\mu_M(n) \tag{4}$$

Multiplying both sides by $\hat{M}/M$ gives the following

$$\mu_M(N)\frac{\hat{M}}{M} = \hat{M}\int_0^N dG(n_e) + \int_\tau W(n,N)d\mu_M(n)\frac{\hat{M}}{M}$$

But this means that the measure μ_M scaled by $\left[\frac{\hat{M}}{M}\right]$ is a new measure that is invariant under the operator $\mathcal{T}$ and entry mass $\hat{M}$. This proves the proposition. ■

D Fixed-point of distribution evolution operator $\mathcal{T}$

Proof. The proof closely follows corollary 4 in Hopenhayn and Prescott [1992]. As shown there, a sufficient condition for the existence of a fixed point is that the distribution evolution operator $\mathcal{T}$ is *increasing*: that is, if $\mu' \succcurlyeq \mu$ then $\mathcal{T}\mu' \succcurlyeq \mathcal{T}\mu$, where $\succcurlyeq$ stands for stochastic dominance.

I proceed as follows. Let $\mathcal{A}$ be an *increasing set*, that is $\mathcal{A}$ equals the set of all elements of the state space that are larger than some element of $\mathcal{A}$. Then, for any $n' > n$, the transition function satisfies $W(n', \mathcal{A}) > W(n, \mathcal{A})$. This is because in the context of the model economy, any increasing set is basically the entire state space beyond a certain point, say n, in which case $W(n, \mathcal{A})$ is the probability that a bank evolves to become larger than that point. But following the intuition behind proposition 2, this probability is higher for larger banks.

Now let f be any increasing, non-negative, and bounded function. Without loss of generality, f can be cast as a measurable indicator function of some increasing set $\mathcal{A}$. This implies that $\int f d[\mathcal{T}\mu'](n) = \mathcal{T}\mu'(\mathcal{A})$. Next, given that in equation (4) in Appendix C the term due to entrants remains unaffected when applying $\mathcal{T}$ on two alternative distributions μ' and μ, we can focus on the term due to incumbents, in particular the transition function W: $\mathcal{T}\mu'(\mathcal{A}) = \int_\tau W(n, \mathcal{A})d\mu'(n)$. Them, since $W(., \mathcal{A})$ is increasing as we showed above, we get the following result, where the first inequality follows from the fact that $\mu' \succcurlyeq \mu$:

$$\int_{\mathcal{T}} W(n,\mathcal{A})d\mu'(n) \geq \int_{\mathcal{T}} W(n,\mathcal{A})d\mu(n) = \mathcal{T}\mu(\mathcal{A}) = \int f d[\mathcal{T}\mu](n)$$

■

E Impact on expected capital growth factor

Proof. The expected capital growth factor (ECGF) is given as:

$$ECGF = \int_{R(1-\chi)} \left(\frac{\psi' - R}{\chi} + R\right) f(\psi')d\psi'$$

The derivative w.r.t. χ (while applying the Liebniz rule) is given as follows:

$$\frac{\partial ECGF}{\partial \chi} = -\int_{R(1-\chi)} \left(\frac{\psi' - R}{\chi^2}\right) f(\psi')d\psi' - \left[\left(\frac{\psi' - R}{\chi^2} + R\right) f(\psi')\right]_{\psi'=R(1-\chi)} (-R)$$

The second term is zero, and the first term is negative since $E[\psi'] > R$, which completes the proof.

■

F Data appendix

The panel dataset consists of US banks that operated between 2000 to 2022. The source is the Consolidated Report of Condition and Income, i.e. Call Reports, obtained via S&P Capital IQ. I delete banks without any data on assets throughout the sample. I also delete observations with negative assets. I further restrict the dataset to commercial banks, savings banks, and savings and loan associations. This results in a total of 10,604 banks. The panel is unbalanced – about half the banks have data for the full time period while others enter or exit the sample during the period.

	Mean	Stdev	P10	P25	P50	P75	P90	N
Assets (USD billions)	1.71	32.07	0.03	0.06	0.14	0.32	0.84	149280
Capital (USD billions)	0.18	3.10	0.00	0.01	0.02	0.03	0.09	140794
Return on assets (ROA) (%): mean	0.76	1.04	-0.02	0.48	0.88	1.26	1.68	149245
Return on assets (ROA) (%): standard deviation	0.72	0.82	0.19	0.26	0.42	0.83	1.66	148820
Dividend payout ratio (%)	4.61	5.70	0.00	0.00	2.81	6.87	11.99	140509

Note: This table reports the summary statistics of the main variables used in the paper. Return on assets is the ratio of net income during a year and the end of year assets. It's standard deviation is computed for each bank over the full sample period period. Dividend payout ratio is the ratio of divided paid during an year to the end of year capital.

I collect annual data on banks' assets, capital, return on assets (net income to assets ratio) and dividend payout ratio (dividends paid to capital ratio). To control for the effect of any large outliers when computing the return on assets and the dividend payout ratio, I winsorize these variables at the 1st and 99th percentiles. The summary statistics of these variables are provided in Table F. Figure 19 shows how return on assets and its standard deviation vary with bank size in the data – this is an important stylized fact used in the calibration of the model.

In addition to the main panel dataset, two other data sources aid the calibration of the model. First is the FDIC's Historical Statistics on Banking that provides information on the frequency and cost of bank failures. Second is the FDIC's disclosures on the insurance premiums paid by depository institutions as a function of the deposit volumes they have. I use these information sources without any data processing.

G Computational details

Solving the bank's problem (Bellman equation): I use Value Function Iteration algorithm to obtain the value and policy functions of the bank's problem. The state-space, which is bank capital, is discretized using 50 log-spaced grid points on the interval

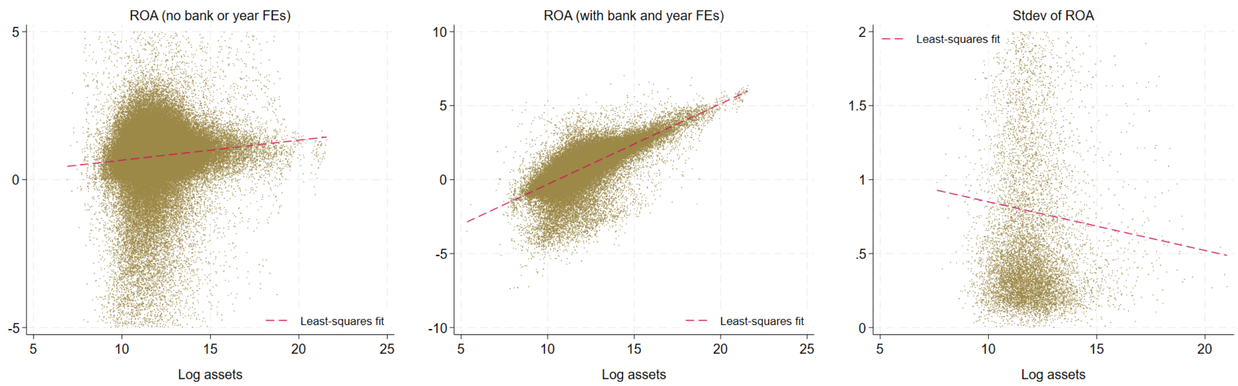

Figure 19: The first panel plots shows the unconditional relationship between banks' return on assets (ROA) and log assets in the pooled data. Each dot represents a bank-year observation (N = 149,245). The second panel shows the same relationship but after controlling for bank and year fixed effects. The third panel shows the link between the standard deviation (stdev) in a bank's ROA during the sample period and its average log assets. In this panel, each dot therefore represents a bank (N = 10,179).

$[\tau, 5000]$ where τ is the smallest possible size of the bank – a bank with capital below this threshold is considered defaulted. Note that the upper limit on the grid is set arbitrarily, and can be adjusted without material consequences for the conclusions in this paper. I use cubic-spline interpolation and linear extrapolation to evaluate the value function at off-grid points. Starting with an initial guess for the value function, I iterate on the solution to the bank's problem subject to the various constraints. I continue to update the value function until the maximum difference (at any grid point) between the old and updated value functions is smaller than a threshold. Once the algorithm converges, I freeze the value and policy functions. Then, for the next steps, I recompute them on a linear grid with 1000 equally spaced points using spline interpolation.

Computing the stationary distribution: To obtain the invariant distribution of bank capital, I construct a state-transition matrix using banks' policy functions. The transition matrix denotes the probability that a bank transitions from one grid point to the neighbourhood of another grid point in the state space. The neighbourhoods are chosen to be of equal size across grid points, so that the entire state space is covered. To handle entry and exit of banks, I introduce a so-called *dump state* in the state-space. The dump-state corresponds to $n < \tau$. A bank whose capital drops below τ (due to low return on assets) defaults and enters this dump state from one of the *incumbent* states. An entrant bank is one that leaves the dump state to enter one of the incumbent states. This approach works precisely because in the stationary equilibrium, the mass of entrants equals the mass of defaulters. The ergodic distribution of the transition matrix gives the stationary distribution, that is the steady-state distribution of bank capital. Essentially, this distribution is one that is invariant when operated upon by the transition matrix. The mass of banks *in* the dump state in the steady-state distribution denotes the mass of defaulting or entrant banks.

Computing aggregates including welfare: Next, I use the stationary distribution of compute overall bank capital, deposits, dividends, and assets in the economy using the set of equations that describe the equilibrium (see Section 4). Using the same set of equations, I then compute total output in the economy, insurance premium receipts, and the shortfall in liabilities of defaulting banks. In turn, this pins down the government budget constraint. Finally, I obtain household consumption and welfare.